AFRL-RH-BR-TR-2008-0067

FATIGUE RESISTANCE ASSESSED IN FIVE TASKS FOR A SINGLE SESSION OF SLEEP DEPRIVATION

Scott R. Chaiken
Donald L. Harville
Richard Harrison

Air Force Research Laboratory

Joe Fischer

General Dynamics

Dion Fisher
Jeff Whitmore

Air Force Research Laboratory

October 2008

Approved for public release; distribution unlimited.

Air Force Research Laboratory
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
Brooks-City-Base, TX 78235

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Human Effectiveness Directorate Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-BR-TR-2008-0067 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//
SCOTT CHAIKEN
Technical Monitor

//SIGNED//
MARK M. HOFFMAN
Deputy Division Chief
Biosciences and Protection Division

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 22-10-2008	**2. REPORT TYPE** Final Technical Report	**3. DATES COVERED *(From - To)*** Feb 2008– Sep 2008

4. TITLE AND SUBTITLE
Fatigue Resistance Assessed in Five Tasks for a Single Session of Sleep Deprivation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Scott Chaiken,[⌂] Donald Harville,[⌂] Richard Harrison,[⌂] Joseph Fischer,* Dion Fisher,[⌂] and Jeff Whitmore[⌂]

5d. PROJECT NUMBER
7757

5e. TASK NUMBER
P9

5f. WORK UNIT NUMBER
18

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

⌂Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
2485 Gillingham Drive
Brooks City-Base, TX
78235

*General Dynamics
Advanced Information Services
5200 Springfield Pike
Dayton, OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Biosciences and Protection Division
Biobehavioral Performance Branch
2485 Gillingham Drive
Brooks City-Base, TX 78235

10. SPONSOR/MONITOR'S ACRONYM(S)
711HPW/RHP, 711HPW/RHPF

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-BR-TR-2008-0067

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited, Public Affairs Case File No. 09-038, 26 January 2009.

13. SUPPLEMENTARY NOTES

14. ABSTRACT To assess whether individuals (*n*=89) could be put on a trait dimension of fatigue resistance (c.f., Von Dongen, Maislin, & Dinges, 2004), we observed performance on cognitive tasks in a single 48-hour sustained-wake protocol. Individual differences in task performance were largest late in the protocol. Next we developed methods for classifying a participant as fatigue resistant or susceptible, as part of a larger project investigating *genetic* factors in fatigue-resistance. We considered a rule based on percent-change decrement with fatigue and another rule based on residuals of task performance predicted by (presumably non-genetic) sleep behaviors, which were shown to bias raw percent-change classifications. Classifications based on ranking residuals were less confounded by sleep behaviors than similar classifications based on percent change. Finally, we assessed the SAFTE fatigue model (Hursh, Redmond, Johnson, Thorne, Belenky, Balkin, Storm, Miller, Eddy, 2004) on sustained performance across the different tasks. While this model is a simulation theory of physiological fatigue, and as such task-independent, we discuss relatively simple ways to put theory and task performances on the same quantitative scale to assess model adequacy.

15. SUBJECT TERMS
SAFTE fatigue model, fatigue-resistance, fatigue susceptible

16. SECURITY CLASSIFICATION OF: Uncl			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Scott Chaiken
a. REPORT Uncl	**b. ABSTRACT** Uncl	**c. THIS PAGE** Uncl	SAR	80	**19b. TELEPHONE NUMBER *(include area code)***

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank

Table of Contents

List of Tables

List of Figures.

ABSTRACT

To assess whether individuals (*n*=89) could be put on a trait dimension of fatigue resistance (c.f., Von Dongen, Maislin, & Dinges, 2004), we observed performance on four cognitive tasks in a single 48-hour sustained-wake protocol. Individual differences in task performance were largest late in the protocol. We developed methods for classifying a participant as fatigue resistant or susceptible, as part of a larger project investigating *genetic* factors in fatigue-resistance. We considered a rule based on percent-change decrement with fatigue and another rule based on residuals of task performance predicted by (presumably non-genetic) sleep behaviors, which were shown to bias raw percent-change classifications. Classifications based on ranking residuals were less confounded by sleep behaviors than similar classifications based on percent change. Finally, we assessed the SAFTE fatigue model (Hursh, Redmond, Johnson, Thorne, Belenky, Balkin, Storm, Miller, Eddy, 2004) on sustained performance across the different tasks. While this model is a simulation theory of physiological fatigue, and as such task-independent, we discuss relatively simple ways to put theory and task performances on the same quantitative scale to assess model adequacy.

According to the National Institute for Occupational Safety and Health (2004) Americans work some of the longest hours in the industrialized world. They also sleep less than they did in the past (National Sleep Foundation, 2005). Consequently, a heavily researched area is the effects of sleep loss on human performance. Extensive data show negative short-term and chronic effects of sleep loss on physiological, attitudinal, and cognitive function (Kryger, Roth, & Dement, 2000; Bonnett, 2000). These effects include mood changes, disorientation, irritability, perceptual distortions, hallucinations, difficulty in concentration, and/or paranoid thinking, depending on the extent of sleep loss. Negative effects have also frequently been shown on cognitive tasks, such as monitoring tasks, speed/accuracy tests, short-term memory, logical reasoning, and mental subtraction/addition (Harrison & Horne, 2000; Pilcher & Huffcutt, 1996).

Fatigue also impairs human performance in a complex manner, indicating the effects of fatigue are multiply determined. For instance, at the level of biology, modal theories of sleep deprivation effects (e.g., Hursh, et. al., 2004) view the sleep-drive homeostatic system in terms of two interactive components, each of which impact performance. One component is based on time spent awake, and is likened to a reservoir depletion process, where “cognitive effectiveness” is depleted. Another is based on circadian rhythms, which vary as a function of time of day, but are anchored or synchronized to the reservoir dynamics by ones habitual sleep/wake times. Both components are set in adaptive opposition, so the circadian process can, to a point, counteract the reservoir depletion process. Between the two processes modal models expect performance to decline during some parts of the day, but also remain stable (albeit lower under sleep deprivation) for long periods of the day (Hursh, et al. 2004).

Nor is it the case that fatigue is always expressed the same way for every kind of task. Complex cognitive tasks that engage more mental resources may fair better (Chee and Choo, 2004). Tasks which involve multi-tasking of conventional cognitive tasks, which individually may be sensitive to fatigue, are not always sensitive to fatigue when they become conjoined (LeDuc, Caldwell, and Ruyak, 2000). Convergent logic-based tasks fair better than divergent innovation-seeking tasks (Harrison and Horne, 2000). Also one may need to take into account the context of measurement, as embedding measurements under operationally realistic conditions can lead to attenuated fatigue effects in those circumstances (Whitmore, Doan, Fischer, French, & Heintz, 2004; Harrison and Horne, 2000).

Finally, and most germane to this study, Von Dongen, Maislin, & Dinges (2004) have pointed out that a sizeable fatigue-resistance trait can be shown using the intra-class correlation. This can be done when participants are observed on the same performance tasks on two sleep-deprivation occasions. For 31 participants, who performed two different tasks, the trait was estimated to be very large for Digit-Symbol substitution (*Z*=.82) and substantial for the Psychomotor Vigilance Task (*Z*=.69). While Von Dongen, et al. (2004) caution that the intra-class correlation can be influenced by the participant's specific aptitude for a task, they also note that the Psychomotor Vigilance Task is a minimally cognitive task, for which healthy, rested individuals, should not be expected to differ in aptitude.

Given individual-differences in fatigue-resistance are likely to be quite substantial, the psychological modeling of fatigue impact is likely to be a complexly-executed issue (e.g., Olofsen, Dinges, and Von Dongen, 2004). A further complexity is fatigue assessment. In particular subjective ratings of sleepiness do not distinguish well between those whose performance change little or much with fatigue (Von Dongen, et al., 2004). Nor have EEG patterns been found to discriminate those whose performance sustains or degrades with fatigue (Galliard, Taillard, Sagaspe, Valtat, Bioulac, & Philip, 2008). For these reasons, fatigue continues to be pragmatically assessed by way of rested and fatigued performance on tasks.

Focus of the Current Study

Here we report on issues closely related to Von Dongen, et al. (2004). The perspective of this study is one of having to identify two different groups of participants that differ maximally on the hypothetical trait of fatigue resistance in preparation for a genetic analysis. Unlike Von Dongen, et al. (2004), we had only one long occasion of testing in which to make the trait classifications on our participants. Given there is a fatigue-resistance trait, most participants would end up (of necessity) in a non-extreme middle range for any reasonable rule we might consider. However, we considered the top 20 and lowest 20, on a preferred fatigue-resistance ranking rule, and then assessed how cleanly that rule separated the extreme groups in the raw data. The general questions we addressed were: what rule to use and how to quality assure a rule's applicability to our goals of follow-on genetic analysis?

Our work is being done in collaboration with the Allegheny Human Performance Laboratory, who will perform whole genome scans on the "amplified" DNA to see whether or not promising gene candidates (identified from animal model studies, Porkka-Heiskanen, 2003; Cirelli, 2002; Tononi & Cirelli, 2001) correlate to a fatigue resistance/susceptibility phenotype. Phenotype is the behavioral profile participants show allowing them to be classified. Further, we hope our phenotypes are attributable to specific genetic differences (e.g., attached to specific brain structure), rather than broad characteristics (e.g., gender, intelligence). Only specific characteristics are likely to afford possible pharmacological interventions to counter fatigue. To increase the probability that our phenotypes have a genetic basis, we investigated how sensitive our classification schemes were to demographics and individual sleep history. Given a source of classification bias (e.g., something plausibly not genetic) could be identified, we developed a general procedure for discounting the bias from our fatigue-resistance rankings.

Finally, as our participant *n* was relatively large for a fatigue study of this protocol-length, we assessed the full-sample results in terms of a representative fatigue model, the Sleep Activity Fatigue Task Effectiveness Model, hereafter, the SAFTE model (Hursh, Redmond, Johnson, Thorne, Belenky, Balkin, Storm, Miller, Eddy, 2004). We chose this model, owing to its availability and our familiarity with its use in risk management. As this was not the primary concern of the study, our (somewhat after-the-fact) execution of this assessment was the best we could do given the data we collected.

SAFTE gives quantitative predictions regarding the physiology that performs cognitive tasks, and provides no predictions of task performance, per se, other than post-hoc regression procedures (see Hursh, et al., 2004, pg. A49). This characteristic is shared by other fatigue models (see Mallis, Mejdal, Nguyen, and. Dinges, 2004, Table 4; see also Gunzelmann, Gluck, Kershner, Van Dongen, & Dinges, 2007). In the course of considering SAFTE we explore simple procedures for putting model prediction and task performance on a meaningful quantitative scale.

Method

Participants

A total of 97 participants were recruited from the local area through e-mail and word of mouth. The participants signed an Informed Consent Document that had been approved by the then Brooks AFB Institutional Review Board, protocol #F-BW-2006-0029-H. Participants included 55 males and 42 females. The mean age of the participants was 26.5 years with a standard deviation of 5.2 years. 62 participants had either recently been in the armed services or were currently serving. Participants were paid $12.50/hour for the 36-hour in-laboratory session, plus a $50 dollar bonus for successful completion of the study. Participants were told they could leave the study at any time via experimenter-paid taxi and would be reimbursed for the hours of their participation (not including training, for which no one was paid).

Tasks and task conditions

Cognitive measures

For this study, we selected three computerized tasks from the Automated Neuropsychological Assessment Metrics (ANAM – Reeves, Winter, Kane, Elsmore, & Bleiberg, 2001), and the Psychomotor Vigilance Task (PVT). The ANAM tasks were delivered on PCs with conventional CRT monitors, mouse, and keyboard, housed on the same kind of ergonomic computer work-desks, while the PVT had its own apparatus. For all computerized tasks, participants sat 50 to 64 centimeters from the screen, with CRT at eye-level.

Each task has a large number of available metrics for which to assess performance. We consider our choices of a “best” metric for each task in the preliminaries of the *Results*. Here we just describe the tasks.

ANAM Math. Participants responded with left or right mouse-clicks according to whether arithmetic expressions (e.g., 6-4-1, 6-2+3, 1+2+1) evaluated to less than 5 (left-click) or greater than 5 (right-click). No expression evaluated to 5. Expressions were in large font; the height of the font being 6% of the screen height and the extent of the string being 23% of the screen width. The task was self-paced, leading to more problems for fast responders. Time out was five seconds and time outs were counted as errors.

Continuous Performance Task (CPT). This is a recognition task using single digit numbers including zero. In a stream of digit presentations (current presentation overwriting the last), participants indicated whether the current digit viewed was the same as the digit preceding the last digit (i.e., same as the one "two back"). Very large font was used (font height 1/5 of screen height; character width about 1/11 of the screen width). The pace of the task was a combination of experimenter and self-paced. If the participant did not respond, a digit would disappear after one second and would "time out" 1.5 seconds after stimulus onset. If the participant responded earlier than the timeout, a new digit would be presented one second later. Time outs were counted as errors.

Grammatical Reasoning. This task presented 48 faceted problems in random order each testing session. In each problem, 3 lines of screen text (font point-size 16) are shown. The first two lines are before-or-after "sentences" (e.g., * BEFORE #, & AFTER #) and the third line is a list of 3 different symbols, i.e., & * #, which is a state of affairs that the two preceding sentences either described correctly or incorrectly. Participants responded with a left-click, if both sentences were true or both false with respect to the third line. If one sentence was true but the other false, a right-click was given. Time out was set at 15 seconds and timeouts were counted as errors.

Psychomotor Vigilance Task (PVT). Participants took this task in the same room as the ANAM. The PVT-192 is a portable, hand-held reaction-time apparatus previously shown to be sensitive to sleep loss (Dinges, Pack, Williams, Gillen, Powell, Ott, Aptowicz, & Pack, 1997). This task randomly and repeatedly delivered a 3-mm visual stimulus to which the participant made a push-button response with the right thumb. The inter-stimulus intervals varied from 2 to 12s. The data extractor for PVT was provided by the vendor. We verified (by looking at raw and summarized data) that some data filtering removed very fast responses (e.g., less than 150 msec) but that no data filtering was applied to very slow responses.

Filler Tasks

We have an independent research agenda attached to two filler tasks, which we will report elsewhere. These tasks were not used to classify our participants as fatigue resistant or susceptible, but they constituted at least half of the testing context and helped keep the participants busy in a work-like fashion during the protocol. We mention anecdotally that several participants reported that it would have been more difficult to complete the protocol without these fillers. So perhaps testing context (e.g., presence vs. absence of work-like activities on clinical fatigue testing) may be an important factor to explore in future studies. Where appropriate we also show how our fatigue classifications impact performance on these filler tasks.

Command, Control, and Communication Simulation, Training and Research System (C3STARS, Tessier, 2006). This was the most important filler task, lasting about 40 min per trial. An earlier version of this task was used to study fatigue in air battle manager trainees (Whitmore, Chaiken, Fischer, Harrison, & Harville, 2007). C3STARS is a conceptually faithful Air Force Command and Control simulation of a mission to destroy enemy Surface-to-Air Missiles (SAMs). Using a point and click (select and command) interface, each participant controlled 3 camera aircraft, two Strike packages (2 bombers, 2 air-to-air fighters, and up to two jammers), and a Tanker. Participant controllable assets had a call sign attached to their icon; whereas other teammates' icons had no call sign. A situation display (radar-scope) allowed friendly assets (blue icons) and hostile threats (red icons) to be monitored and manipulated in real time via switch action buttons, communication windows, and information windows.

Half of our participants played C3STARS as part of a team of three in a joint war. The other half played C3STARS in a solo-player mode, which fought a war scaled to be 1/3 the size of the team war. Both conditions played in separate rooms of similar dimensions and with similar testing-station layout (i.e., each player facing their respective CRT and respective room wall). Whenever available the solo-condition had 3 players to match the 3 necessary players of the team condition. Given a team-condition participant decided to leave the study, the condition changed to the individual condition and their C3STARS data were discarded.

For the team condition, the region of responsibility was an entire geography. Teams were told (truthfully) that the enemy targets occurred randomly in that geography with SAMs clumping together more in some areas than others. Trainers suggested that teammates coordinate targets among themselves in order to balance the workload, and teammates communicated with each other either verbally or by text messaging to do this. For the solo condition, a player executed the war in one of 3 possible lanes (a balanced factor across individuals) dividing the whole scope; however, as C3STARS allowed panning and zooming of the situation display, the size of the geography in the situation display can be at the prerogative of the player. The un-played lanes had enemy activity and other (passive) friendly assets in them. Solo condition participants were taught to ignore the other lanes by going up the center of their played lane. Communication between solo players was not allowed.

Both the team and individual C3STARS conditions had twelve kernel scenarios, twelve 120-degree rotations of the kernels, and twelve 240-degrees rotations of the kernels. The 12 scenarios had random factors determined by their seed number 1 - 12, with a team seed of 1 being a scenario that is unrelated to a solo-condition seed of 1 (given the number of random events differed between a solo and team scenario). Within a condition, the seed numbers governed the placement of ground targets within enemy territory, launch times and flight plans of enemy fighters, and flight plans for friendly aircraft prior to their assignment to missions. Within team and individual conditions, scenarios were equivalent on number and kind of enemy threat and assets to cope with them.

Testing sequences used odd-numbered scenarios that changed seed number every trial and rotation-type most every trial (whereas training sequences were even scenarios). Testing sequences were rotated across participants during the course of the study, so that the same seed numbers would be used both early and late in the protocol across different participants and across the different conditions.

C3STARS activity came in roughly sequential phases that could overlap. We give two figures in an appendix to further give a sense of the depth of the game. Figure A1 is a time chart of the various phased activities in C3STARS and Figure A2 is a screen shot.

During the pre-mission planning phase (5 minutes) simulation time was stopped, and participants set their (Intelligence, Surveillance, Reconnaissance) ISR assets on reasonable default routes into enemy territory with way-point missions. They also gave rendezvous orders for their first Strike package of in-flight aircraft, consisting of two fighters, two bombers, and a lead jammer. Friendly bombers had two bombs, friendly fighters had two superior missiles (relative to the enemy) and two inferior missiles, and jammers had no weapons. Individuals (in either team or individual) conditions had 1 extra bomb and 1 extra premium missile relative to the amount needed to perform a perfect mission. Enemy opposition per individual player (i.e., multiply by 3 for per team) was six real SAM sites (with one SAM requiring 2 bombs), seven hostile air-to-air fighters, and seven SAM decoys.

After simulation time started, small sensor-blips appeared on the screen in enemy territory, indicating something on the ground. The participants “tagged” minute blips (less than ¼” height), which looked like an upside-down T. Tagging required a psychomotor sequence of clicking the lower-left-hand side of the blip; if successful, a circle appeared around the blip and pressing ENTER (before the circle disappeared, in 1 to 3 seconds) finished the action. A tagged item resulted in a call-sign (e.g., SA6C) and a red icon being attached to the sensor blip. The icon indicated the nature of the target according to how many ISR assets saw it after tagging. With no ISR asset seeing it, the icon signified an unknown; with one asset, the icon signified a best guess (e.g., a real SAM or a harmless decoy with 67% probability), and with two assets, the icon indicated ground truth. Without tagging the item, participants could take no action against it; however, the item could still kill given it corresponded to a real SAM.

After tagging, participants adjusted their ISR mission routes to their more informed perspective of the enemy theatre. They also worked on forming their second package (two each of fighters, bombers, and jammers) which finished launching from a rear, friendly, air base, about 3 minutes after the simulation started. Participants had to refuel base-launched fighters before rendezvousing them with the rest of the package; otherwise the fighters were lost to fuel outages.

After their first Strike package rendezvoused outside of enemy territory, participants proceeded to bomb fully-identified SAM sites. Jammers blinded the enemy SAMs; however, hostile air-to-air fighters, launched periodically (all seven launched within 15 minutes at regular intervals), were *not* blinded. Friendly fighters from the package protected the rest of the package from these hostile fighters. ISR assets were generally safe from SAMs and hostile fighters by flying outside of the altitude band SAMs and hostile fighters could shoot. Similarly, all other friendly aircraft were constrained to fly within the sensitive band.

Participants had to manage their assets so correct targets were bombed and enemy fighters killed, while remaining fueled and not straying too far from their jammer. What makes the task challenging is that any given mission takes assets away from the jammer making them susceptible to SAMs. Participants therefore had to ensure assets come back to the jammer after a short mission or insure the jammer stayed abreast of all their package assets. Further, if the jammer inadvertently led too much, it was shot down by hostile aircraft, at which point the entire package would usually be lost to SAMs. Monitoring the assets for risk during mission execution is a highly dynamic and continual activity.

Hostile fighter aircraft had a simple intelligent agent behind them. They flew their randomly scripted surveillance missions, but if their radar (whose detection range just exceeded their missile range) saw a friendly, they left their mission and intercepted the friendly. Hostiles that killed a friendly returned to their base (a fact participants were trained to expect). Hostile SAMs killed multiple friendlies, given such friendlies had no jammer protection and had been tracked for 30 seconds.

Assuming a SAM was bombed, there was a "paper-work" trail of three steps to drop the SAM from the scope and receive full credit. This involved: 1) requesting a battle-damage-assessment (BDA) mission on it (resulting in the SAM's icon becoming annotated with "RDA", for request damage assessment) 2) committing two ISR assets to get snapshots (requiring maneuvering two ISR assets within closer range of the SAM), 3) receiving two "Target destroyed" snapshot messages, from each of the ISR assets (requiring monitoring the communications window or recognizing the synthetic speech of their asset over headphones), and finally 4) dropping the SAM from the scope (requiring selecting the correct SAM icon and performing a drop switch action). These paperwork actions could be done without having actually shot or destroyed the SAM; however step 3) was illegal without a prior RDA (as communicated by a text/speech message). Participants were trained to expect one out of every six SAMs to require two bombs for a sure kill (i.e., BDA snapshot messages come back "Target survives" after the first shot for these).

C3STARS mission outcome was defined as number of hostiles shot (hostile fighters and SAMs) plus number of correctly completed battle damage assessments on the SAMs ***minus*** the total number of friendlies lost, either to enemy fire or running out of fuel. Mission outcome relevant scores (i.e., pie charts showing components of mission outcome performance) were shown participants after scenario play. Teams were required to look at a single pie chart for their team together; solo players were allowed to show each other their pie charts at their option.

Synthetic Work (SYNWIN version 1.2.37). SYNWIN (formerly SynWork, Elsmore, 1994) is a simulated work environment that required the participant to monitor four different activities each occurring in one of four quadrants of the computer screen. Some of the tasks have auditory alerts, so participants wore headphones for SYNWIN. Response was via a mouse. A continuously updating cumulative work score was shown center screen. Correct responding gained points; incorrect responding decreased points by the same amount. Time outs lost points on three of the four tasks (the fourth being self-paced with no time out). SYNWIN was set to last 10 minutes, and we scored it using the total number of work points earned by that time.

In the upper-right quadrant of SYNWIN was a four-digit two-number addition task (the self-paced task). The participant controlled which number appeared in a results slot (i.e., ones, tens, hundreds, and thousands) by clicking a plus-button or minus-button underneath the slot. When all four slots were filled as desired, the participant clicked a nearby “done” button.

In the upper-left quadrant was a Memory-Search Task, using a six-letter memory set for the entire 10 minute period and a left or right mouse-click to signal letter-present or letter-absent. After a short study time (10 seconds), the list disappeared but could be made available for re-study when its empty field was clicked. Re-study had an error penalty attached. After an initial probe that had a 20-second time out, subsequent probes of the list changed every eight seconds.

In the lower-left quadrant is a Fuel-gauge task. A fuel gauge line moves from right to left, traversing green, yellow, and red zones of the gauge. Clicking on the fuel-gauge restarts the process. The closer one restarts from the red zone, the more points are awarded. If the fuel gauge reaches zero, an auditory cue alerts the participant, and 10 points are lost for each second the fuel remains at zero.

In the lower-right quadrant is a high-tone detection task, for which a big red alert button is shown. When a high tone is detected the participant has five seconds to click that button to get points for successful signal detection. Lower tones are given with greater frequency, and responses to these are errors.

Questionnaires

Recent Sleep Activity Log. Each participant was provided with a single sheet of paper on which to record their wake and sleep times for the four days (Tuesday through Friday) preceding the experimental session. The form was broken into 48 half-hour intervals for each day. Participants filled in the intervals as needed.

Demographic survey. This was a short-answer survey containing questions regarding the participant's age, gender, tobacco use, education level, and military career experience. This survey was augmented after the first 24 participants to query video-game enjoyment/experience. We did this as some participants reported liking, and/or doing well at SYNWIN and C3STARS because they liked video games.

Sleep Behavior Questionnaire. Participants answered a 12-item questionnaire describing their habitual sleep behavior. It covered whether or not they were a shift worker, and how long their typical sleep periods were, considered separately for both weekdays and weekends.

Description of Experiment

Training

Participants were required to participate in a four-hour training session on each of two consecutive days prior to the experimental session. Training started with signing the Informed Consent Document and a short brief from the medical monitor for that weekend run, which focused on who should not participate in the study based on prior medical conditions.

Participants (12 maximum and 7 minimum per cohort) were assigned to one of two "time streams," which continued through testing. Streams are defined by whether C3STARS or ANAM lead the testing block. Time streams were required because rooms and computer equipment were not sufficient to test everyone on the same tests at the same time. Therefore, each training session was broken into four one-hour segments with participants rotating between the C3STARS rooms (i.e., a team and an individual room, each holding three people) and the single ANAM testing room (holding six people). Table 1 describes the training events for the two streams.

Table 1

Training Schedule

	Day-1		Day-2	
Time	Stream 1	Stream 2	Stream 1	Stream 2
1800	C3STARS	SURVEYS ANAM (2X)	C3STARS (2X)	ANAM (5X)
1900	SURVEYS ANAM (2X)	C3STARS	ANAM (5X)	C3STARS (2X)
2000	C3STARS	ANAM (4X)	C3STARS (2X)	PVT, ANAM SYNWIN (2X)
2100	ANAM (4X)	C3STARS	PVT, ANAM SYNWIN (2X)	C3STARS (2X)

Notes.

SURVEYS are sleep behavior and demographic questionnaires. ANAM refers to ANAM Mathematical Processing, Grammatical Reasoning, and Continuous Performance Task, given in that order. PVT is the Psychomotor Vigilance Task. C3STARS is a command and control battle simulation and SYNWIN is a synthetic work environment.

A standard training regimen, divided over two nights, was devised and given to all subjects by two subject-matter-experts for C3STARS. The C3STARS trainers alternated training roles (i.e., individual or team) each testing occasion. The first night/first hour involved in depth discussion of buttons and controls (in the context of tactics) using the pause facility of the simulation to keep trainees synchronized. This was followed by two abbreviated plays of the same session. The second night involved four abbreviated (30 min) plays that were monitored and coached. Participant learning varied and for the majority of participants C3STARS performance was pre-asymptotic by the time the protocol started.

ANAM training began with a discussion on proper task-taking procedures (e.g., posture and mouse-positioning). Following this was an introduction to the three ANAM tasks, with the optimal strategy for learning each task given through the on-screen instructions and through the ANAM trainer. As a general strategy, participants were told to concentrate on accuracy first and speed later. Participants recorded their accuracy and speed (given by the ANAM at the end of a task block) to allow the trainer to assess how well asymptote was approached. Over the course of the 3.5 hours, participants would complete at least ten cycles (15-20 mins) and most were asymptotic.

The PVT was presented to the participants in the cognitive testing room on the second half of the second night of training, just prior to an ANAM practice. Participants were instructed on the proper way to take the PVT. This included sitting forward with both hands on the PVT box, elbows on knees, and keeping the right thumb on the response button. They then completed a one-minute demonstration of the actual task. PVT was trained by the same person who trained the ANAM.

SYNWIN was introduced in the cognitive testing room at the end of the second day of training, just after an ANAM practice. A total of two 10-minute training blocks were administered. However, the first training block involved instructions and a 2.5 min practice for each of the four sub-tasks. The second block was an administration of the task as it was tested (all sub-tasks together). On screen instructions gave participants an optimal strategy for learning the task, which involved concentration on two of the tasks initially (i.e., math-problems and fuel-gauge monitoring), with expansion to the other tasks as performance on the initial pair of tasks improved.

Testing

Participants arrived at 1800 on Friday evening to begin 36 hours of in-laboratory sleep deprivation which would follow an estimated 12 hours of pre-laboratory wake time. The first hour was a working dinner, during which participants filled out other exploratory questionnaires, and turned in their recent sleep logs. This was followed by 3 hours of testing, to make up one four-hour testing block, which would be repeated eight more times. Breaks included meals at appropriate times. Table 2 shows the events of one testing block.[1] Three proctors (at least one female) continuously monitored and managed the participant flow along the protocol's scheduled events.

Table 2

A Four-Hour Testing Block

Time	Stream 1	Stream 2
2200	HOUR BREAK	HOUR BREAK
2300	C3STARS	PVT ANAM
2345	ANAM PVT	C3STARS
0030	C3STARS	SYNWIN ANAM
0115	SYNWIN ANAM	C3STARS

Notes.

Tasks identified in Table 1. After the hour break comes four 45-min task blocks.

Breaks were held in a common lounge area of two large rooms with tables, televisions, and video-gaming systems. Participants stayed within the building throughout the protocol, and were allowed to take showers during the breaks. Participants finishing their tests early read, played computer solitaire, or conversed in their testing room (provided the entire room had finished), until it was time for their next rotation or the next break. Each testing room had a bathroom.

DNA sample collection was accomplished via buccal cells collected with Scope® mouthwash on the first evening of the fatigue protocol at about 2200. Participants were asked to wait at least one hour after eating or drinking before providing buccal cells. Participants swished 10 ml of the solution 10 to 20 times before expectorating into a 50 ml tube, labeled with their anonymous subject number. The tubes were sent by express mail at room temperature to the Center for Genomic Sciences, Allegheny Singer Research Institute.

After the completion of the ninth testing block, subjects were released to go home. Participants left the study via study-paid taxi or via a pre-arranged pick up by a family member or friend.

Results

Missing Data

Analyses are on the 89 participants who completed the study (out of 97 who started the study). Seven participants did not complete the protocol and one participant was considered inappropriate to analyze owing to night-shift work the week of the protocol.

For the remaining participants, missing data were an infrequent occurrence owing to various equipment/software reasons. Our general strategy will be to report basic analyses (e.g., general fatigue effects) with complete data only. For our classification procedures, which select a sub-sample of people who are most fatigue susceptible or resistant, simple data estimation techniques, such as using the maximum or average on a set of *available* scores, allowed us to be more complete in our data usage.

Task Metrics

Prior to any classifications, we decided on a single metric for each task, which would best summarize participant performance on a given testing trial, derived from vendor-provided measures. For ANAM Math Processing (abbreviated as ANAM Math), participants received more problems the faster they responded. The least controversial performance measure, under these circumstances, is “right minus wrong,” which weights both speed and accuracy (and corrects for guessing where chance is at 50% accuracy). This metric was found to be superior to the vendor-provided metric “throughput” as participants could be found that did well on throughput but poorly on accuracy (i.e., near chance), when fatigue levels were high.

For ANAM Grammatical Reasoning (abbreviated as ANAM Gram), every participant received 48 problems. Speed wasn’t emphasized in the instructions and there was no reason to account for differences in participant speed. Therefore, accuracy was chosen as the best metric for this task.

For the Continuous Performance Task (abbreviated as ANAM CPT), list items are both participant and computer paced. While it is true that a variable number of problems could be given with differences in response rate, this characteristic was observed to be less variable than it was for ANAM Math. Also as ANAM CPT could be more of a test of working memory, we used accuracy as a best metric, as this is typically the measure of choice for working memory tasks (e.g., Kyllonen & Christal, 1990; Woltz, 1988).

Von Dongen, Baynard, Maislin, Dinges (2004) reported that the Psychomotor Vigilance Task (abbreviated as PVT) loaded a separate "factor" from the factor their selected cognitive tasks loaded. This was found in an orthogonal, varimax factor analysis of fatigue-impairment scores for cognitive tasks and PVT. As the PVT is a sustained attention task (e.g., vigilance), which may show distinct properties from cognitive processing, we decided to consider PVT as a basis for a single-task rule for classifying people as resistant or susceptible.

For the PVT, one frequently finds both mean reciprocal response time (e.g., Dorrian, Roach, Fletcher, Dawson, 2007) and number of "lapses" (i.e., number of RTs exceeding 500 msecs, e.g., Van Dongen, 2006) as the reported performance metric. However, our sample also had substantial occurrences of "false starts" in some subjects. These are responses in the interval after a valid stimulus is responded to but before a next valid stimulus has been presented. Table 3 shows the analyses for all the possible PVT metrics we considered.

Table 3

Fatigue impact (means and standard deviations) on Psychomotor Vigilance Task (PVT) for different performance metrics

Trial	Lapses[a]	sd	Falses[b]	sd	PVT(Rate)[c]	sd	PVT(errors)[d]	sd
1	1.2	1.7	3.1	5.3	4.1	.44	4.4	6.0
2	2.3	3.7	3.1	4.6	3.9	.52	5.4	6.7
3	7.3	8.6	2.7	4.2	3.4	.69	10.0	11.0
4	17.5	13.8	5.1	7.3	2.8	.73	22.7	17.9
5	14.8	11.2	6.6	10.8	3.0	.77	21.4	17.0
6	16.3	10.4	6.2	11.1	2.9	.68	22.6	16.5
7	13.2	10.7	5.2	7.2	3.1	.70	18.4	13.3
8	14.0	10.1	6.5	12.1	3.0	.67	20.4	16.0
9	20.9	14.6	7.8	16.7	2.6	.77	28.8	23.4
F[e]	62.2		5.65		113.6		49.8	

[a] *number of responses with reaction time greater than 500 msecs*

[b] *number of responses without a valid stimulus (false starts)*

[c] *mean reciprocal response time to a valid stimulus*

[d] *PVT(errors) is lapses + false starts; MSE(6205, 125)*

[e] *F-values for Trial main effects tests: n = 86; df(8,680); p < .001.*

As both false starts and lapses significantly increase with fatigue, we considered a "lapses + false starts" metric as preferable to simply measuring PVT on lapses alone. We will abbreviate "0-(PVT lapses + false starts)" as PVT(errors), to express the idea that both kinds of responses are errors with respect to the task. We subtract the total number of errors from zero to make the PVT(errors) scale consistent with the other performance scales, where higher score means better performance. PVT(errors) is not the strongest metric to consider against fatigue. PVT(rate), or PVT mean reciprocal response time, had a considerably larger *F* for the time (i.e., fatigue) effect. However, the problem with using it in our study is that it does not consider false-start behavior (admittedly a smaller *F*) as part of the degradation with fatigue. Participants with many false starts after a sustained wake are likely to show more rapid regular responding just because they have false starts. This would be a kind of reaction time "guessing."

Because the PVT has historical importance to the fatigue literature, was being considered as a "single-task" rule, and, finally, has uncertainty as to the best performance metric to characterize fatigue impact (both in the literature and our study), we scrutinized the metric choice for PVT in further analyses in these results. At the end of this process, we give arguments favoring the use of error counts as the preferred metric (see Discussion).

Basic Fatigue results

Before describing how performance on ANAM and PVT can be used to classify people, general results relating to fatigue are shown in Table 4. These are the results of repeated measures ANOVAs for all the individual-level tasks of the study (excluding PVT which is shown in detail in Table 3). As expected, ANAM shows robust time (fatigue) effects. Figure 1 shows fatigue functions for the ANAM tasks and PVT. PVT was tested 9 times, the others tested 18.

Table 4

Fatigue impacts on tasks

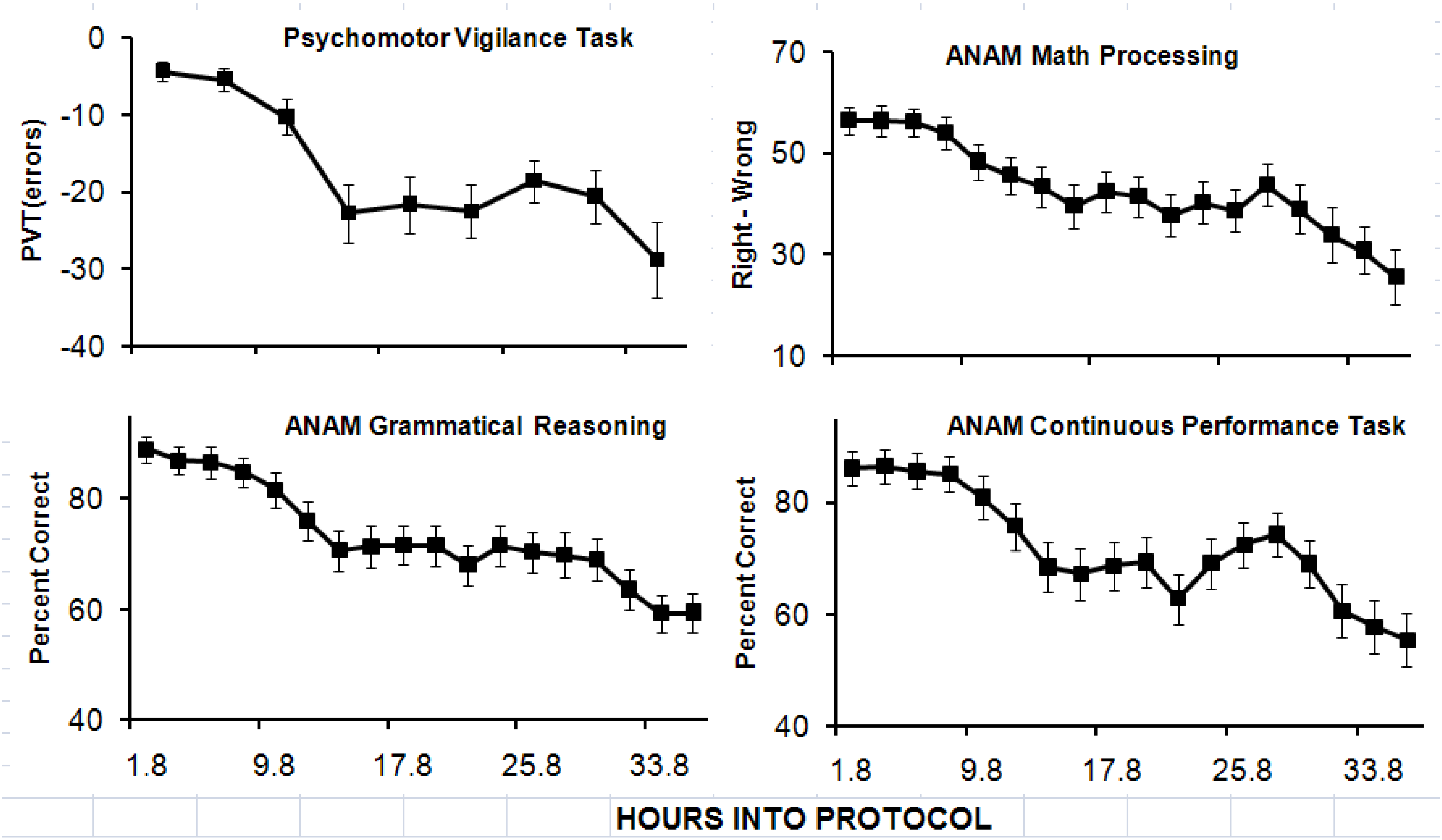
Task	n	*df* /*MSE*s[a]	*F* for time effect[*]
Math Processing	87	(17,1462) / 6695, 164	40.8
Grammatical Reasoning	85	(17,1428) / 7080, 127	55.8
Continuous Performance Task	86	(17,1445) / 8362, 175	47.7
Synthetic Work for Windows	88	(8,696) / 6758508, 337727	20.0
C3STARS Individual	44	(17,714)/ 90, 28	3.19

Note. Fatigue is assessed by trial number.

**Fs are significant at p < .001, or smaller.*

Figure 1 Caption. Fatigue plots of mean performance (raw data) against hour into the protocol. X-axis ticks reflect the end of a Time Stream 1 testing segment and the beginning of a Time Stream 2 testing segment (see Table 2), leading to an approximate testing time (i.e., plus or minus 45 minutes as a maximum error). PVT(errors) is a total number of errors (long-responding or false-responding) subtracted from zero to make PVT's scale consistent with other tasks (i.e., low scores are poor performance). Error bars center 4 std. errors on their mean.

FIGURE 1



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

Task Correlations

Table 5 shows performance correlations between the individual-level tasks, both at the beginning of the protocol (upper diagonal matrix) and at the end (lower diagonal). We considered the average of the first four protocol scores for the ANAM tasks (the average of the first two for PVT) as early performance and the average of the last 4 protocol scores (average of the last two for PVT) as late performance. We correlated ranked performance (e.g., subject standing out of an *n* of 89 on a task) rather than raw performance to remove any effects of outliers (in the raw metric) on the correlations. Using raw scores would result in similar, but slightly lower correlations, than those shown in Table 5. Table 5 clearly shows correlations between tasks as generally higher for late rather than early performance. This is consistent with the idea that individual differences in task performance increased with fatigue.

Table 5

Ranked early performance correlated between tasks (upper triangular matrix) and ranked late performance correlated between tasks (lower triangular matrix)

	1.	2.	3.	4.	5.	6.	7.
1. Math	--	.49	.42	.20 ^{ns}	.33	.42	.15 ^{ns}
2. Gram	.70	--	.59	.36†	.28	.53	.03 ^{ns}
3. CPT	.80	.72	--	.26 ^{ns}	.44	.56	.13 ^{ns}
4. PVT(errors)	.54	.50	.68	--	.13^{ns}	.22 ^{ns}	.13 ^{ns}
5. SYNWIN	.67	.64	.70	.50	--	.60	.12 ^{ns}
6. C3STARS_Ind	.47	.45	.39	.21 ^{ns}	.58	--	.08 ^{ns}
7. PVT(rate)	.50	.51	.64	.72	.49	.11 ^{ns}	--

Note. Ranked early performance is the average of the first 4 protocol trials for ANAM (2 for PVT); ranked late performance is the average of the last 4 protocol trials for ANAM (2 for PVT). n1,n2=89, 89 for ANAM Math Processing, Grammatical Reasoning, and Continuous Performance Task; n1,n2=87,86 for Psychomotor Vigilance Task; n1,n2=88,86 for SYNWIN; n1,n2=89,89 for PVT; n1,n2=46,46 for C3Stars Individual Task.

^{ns}not significant, † p<.05 two-tailed; no superscript: p < .01, two-tailed.

Fatigue results related to a fatigue theory (part 1)

Given the different scales for each task in Figure 1, one might wonder if each task measures fatigue impact to the same degree, and as a related matter, how well the fatigue impacts shown by these tasks compare to a theoretical model's prediction of fatigue impact. We used an "intra-subject z scaling" transformation applied to task performance to address the first question and predictions of SAFTE (Hursh, et. al., 2004) to address the second question.

The intra-subject z-scale is a way of putting each task on a similar scale of fatigue impact. It does this by expressing each participant's within-protocol task performance in terms of their performance variability on that task during the protocol. If a participant contributes 18 scores for a particular task, re-scaling his or her scores as "intra-subject z-scores" involves subtracting the mean of their particular scores over the protocol from each of their scores and then dividing that by the standard deviation of their original raw scores. This kind of transformation changes the expression of individual differences (i.e., the ranking of fatigue effects between subjects) quite dramatically, so we would never use this for classifying people according to the size of their fatigue effects. However, we use this approach for comparing fatigue impact *across different situations*, and as a kind of model-fitting (treating the model prediction as a kind of "situation"). This approach is *not* model-fitting in the specific sense of our having a model for each of the tasks we observed under fatigue. We will defer a defense of this procedure as "model-fitting" for the Discussion.

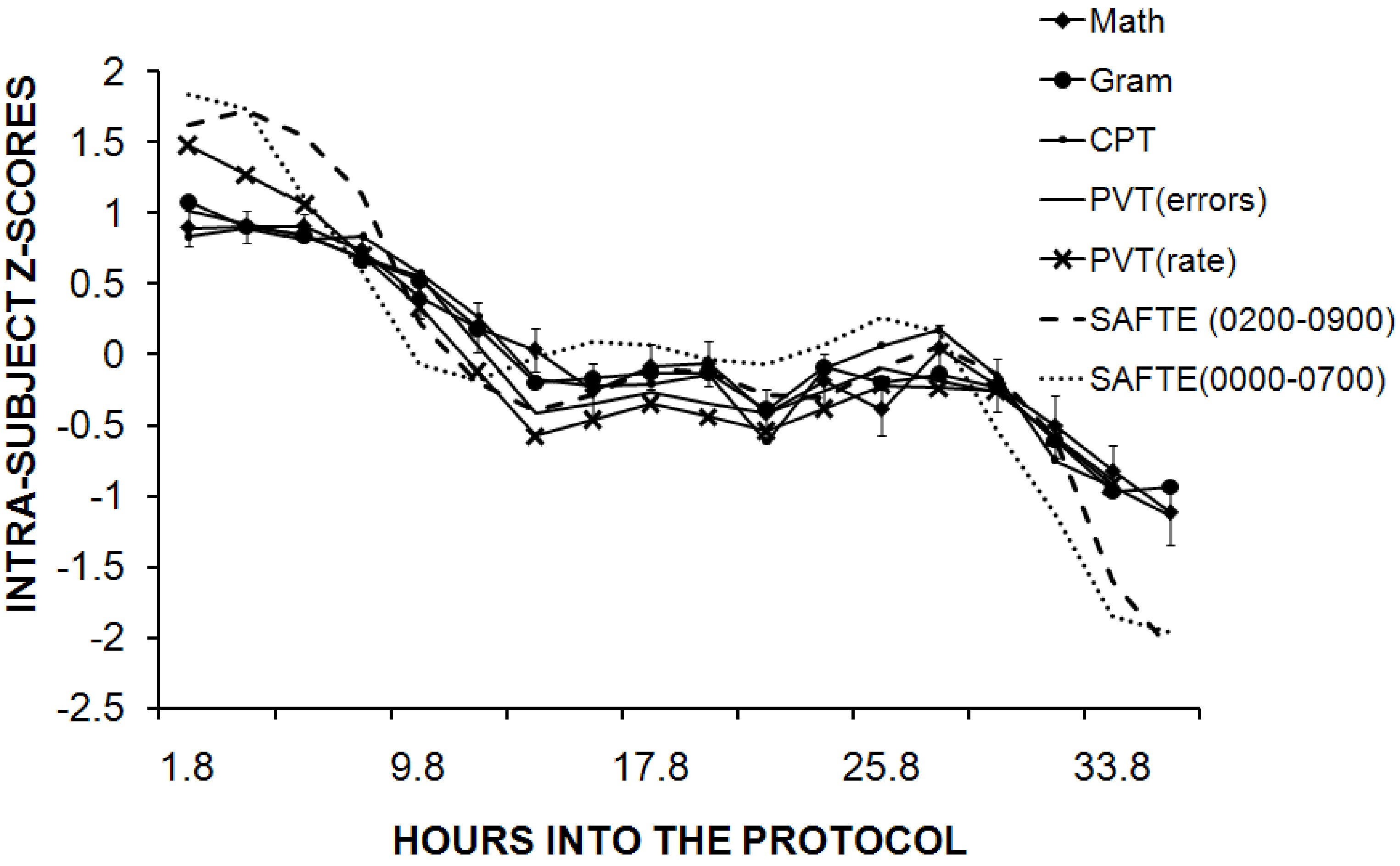
Figure 2 shows a crowded display of results, but the take home messages are fairly simple, once the different types of functions in the figure are defined. All the solid-line functions are observed task-performance functions-- i.e., averages over the individual intra-subject z functions for a particular task. There are two observed functions for the PVT, one for PVT(errors) metric and one for PVT(rate) metric. For the PVT functions in Figure 2, only the odd trial points reflect data; even ones are interpolated to allow our plotting package to show all tasks on the same plot.

PVT(rate) diverges from PVT(errors), and the other tasks, by showing a steeper fatigue decline originating from a higher starting point (i.e., a paired *t-test* yields $t(88) = -9.12$, $p < .001$, for comparing the PVT metrics on the first point). There are significant differences at later points, but none are as large as the first point, and the largest later difference is in the opposite direction from the first point (i.e., at 9.8 hours into protocol, $t(88) = 3.65$, $p < .001$). The PVT metric scores correlated least when fatigue was the least and correlated most late in the protocol. For example, the correlations between the metric scores for PVT trials 1, 2, and 3 were .30, .32, and .53 respectively ($n = 89$, $p < .01$, two tailed); for PVT trials 7, 8, 9, the correlations were .77, .72, and .63 respectively (n = 89, 87, 88, $p<.001$, two-tailed). We think this information may provide another argument (independent of the "false starts" issues) for preferring an "error-based" PVT metric over a speed based metric. We take this up more in the Discussion.

ANAM and PVT(errors) performance seem to fatigue similarly. This is not to say there is insufficient power to detect differences between some of these tasks[2], but only that under this transform, and in a zoomed-out perspective, the majority of tasks reflect fatigue impact similarly. To help gauge variability of the family of tasks observed, error bars are shown on each point for ANAM Math, reflecting a total range of 4 standard errors of the mean (based on the observed *sd* at a given ANAM math point in the intra-subject scale).

Figure 2 Caption. Fatigue plots, attained from averaging within-individual z-transformed performance functions for multiple tasks compared to SAFTE fatigue model predictions on the same scale. For PVT (errors and rate) only odd numbered points reflect data. Even points are interpolated to allow PVT (administered 9 times) to be plotted with ANAM tasks (administered 18 times). Representative error bars (in the manner of Figure 1) are shown for ANAM Math for the entire protocol.

FIGURE 2



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

Also plotted in Figure 2 are two SAFTE "cognitive effectiveness" predictions, re-scaled to intra-subject *z* units (i.e., as one would rescale an individual's observed fatigue function, prior to finding the average over all participants). SAFTE predictions are the broken lines with no markers. The dashed line is a function that fits best. It was found by comparing *z*-ed SAFTE predictions for different possible sleep-behavior profiles and choosing the profile with the minimum squared-deviation from the observed functions. The best-fitting profile we observed corresponds to individuals sleeping 7 hours habitually, which is what our participants report, and habitually going to sleep at 0200 and waking at 0900. The sleep and wake times of the best-fitting profile is about 2 hours later than participants reported for their sleep behavior the week of the study. Unfortunately, we did not ask participants to directly estimate their habitual sleeping and waking-up times in our sleep behavior questionnaires (as we did for habitual sleep amounts). Therefore another SAFTE prediction, corresponding to their recent sleep behavior, is shown as a dotted line (i.e., the profile of 7 hours habitual sleep with 0000-0700 sleep/wake times). While we don't know which SAFTE profile best reflects our participants, both SAFTE predictions "fatigue" roughly as the tasks do (i.e., decline, long-plateau, decline). Both predictions also indicate significant lack of fit at the beginning of the protocol and significant lack of fit at the end of the protocol. However, this is not the most appropriate way to use a z-transform on observed data to assess SAFTE's prediction. We return to this issue in the Discussion.

Fatigue classification rules

The 48-hour termination time for our study was determined by our research partners' desires for: 1) a challenging sleep-deprivation interval and 2) a protocol that ended when both the circadian and reservoir-depletion effects of fatigue would be expected to be maximal (i.e., early morning release). Because of these considerations, and the fact of a long plateau region in the middle of the protocol, the rules we considered focused on performance at the beginning and end of the protocol and excluded the data in the middle. For each rule considered, we use the same definition of baseline (beginning) and fatigue endpoint (end of protocol). The baseline was the maximum of the first four trial scores if an ANAM task, or the maximum of the first two if the PVT task. A maximum would select the best score, given participants were fatiguing rapidly or still learning over the first trials. For fatigue-endpoint we used the average of the last four trials if ANAM, and the average of the last two trials if the PVT task. Using the constraints just described, four candidate rules were considered. One rule turned out to be highly similar to another, and another rule was deemed inferior because of not providing a continuous fatigue-resistance grade. Therefore, we settled on two rules described below as the best ones to further assess.

Percent-change rule

This rule creates a score for each task via the formula (endpoint-baseline)/baseline, which is a metric used in the fatigue literature (e.g., Caldwell, Mu, Smith, Mishory, Caldwell, Peters, Brown, & George, 2005). Given an appreciable effect of fatigue, the endpoint is expected to be less than the baseline. Therefore, percent changes that are more negative show more fatigue effect. For ANAM, the participant’s percent-change is found for each task and that percent change is ranked relative to other participants on the same task. Participants high in rank (i.e., numbers closer to 89 than to 1) are participants high in fatigue resistance. A participant’s average percent-change rank over the three ANAM tasks is then found and used to classify participants. For notional cut points we decided to aim for about 20 in each of our extreme groups. Because of ties we settled for 21 in a group for some rules and group categories. For PVT, simple change (endpoint-baseline) was used to rank participants, as zero is an allowable and frequent baseline score, making any transformation to percent-change problematic. In summary, the percent-change rule for ANAM yielded 20 susceptibles and 21 resistors (and an unclassifiable middle group of 48). For PVT, a simple-change rule also yielded 20 susceptibles and 21 resistors (and an unclassifiable group of 45).

Residual-score rule

Given our definition of fatigue baseline and fatigue endpoint, we predicted endpoint performance using baseline performance on the same task (i.e., simple linear regression, n = 89). The participant-specific residual of this regression is a kind of fatigue-resistance score (see Woltz, 1988, for an example and some discussion of this technique applied to a different kind of cognitive-trait measure). This residual-score is an index of fatigued performance not accountable by, or controlling for, initial (non-fatigued) performance. Such residual-scores can be standardized (i.e., a z-residual) and saving these, as a new score, is easily accomplished using a SPSS^{tm} regression option. As with ranks in the percent-change rule, standardized residual-scores from different tasks can be averaged together without any one task dominating the variance.

Another boon for residual scores is that they readily extend their definition to other covariates. If one knows that some hypothetically non-genetic participant characteristic, such as average weekday sleep, significantly relates to their percent-change (i.e., fatigue effect) on tasks, as we discuss later, one can add the problematic confound to the regression equation along with the baseline score, and have both predict the endpoint score. The residual score, from *that* *regression*, will be statistically independent of both initial performance and the confound. One can use rankings of *these residuals* as a “corrected” ranking for fatigue-susceptibility. For example, if someone appears to be highly fatigue susceptible in the raw data, but also reports sleeping only five hours on weeknights, the regression/residualization procedure discounts that person’s susceptibility ranking (as might have occurred in the raw data) by removing that part of his or her score owing to the impact of amount of weekday sleep, as inferred from the relation of weekday sleep to the sample’s performance. Similarly, if a participant sleeps five hours a night weekdays but shows the average amount of performance change with fatigue, correcting the ranks can lead to that person being considered fatigue-resistant.

We created an ANAM-based residual-score classification in the manner of the percent-change rule. We averaged z-residuals for the 3 ANAM tasks and took the top 20 and bottom 20 residuals as our classification (in this case there were no ties at group boundaries). We computed a separate score for the PVT. As suggested above, we obtained *the final set of residual scores* by correcting for sleep behavior variables that significantly related to *uncorrected percent-change-rule fatigue classifications*. This model can be thought of as a percent-change classification rule corrected for a bias model. The factors in the bias model can be easily described. They are an amount of weekday sleep participants report getting habitually and their going-to-sleep and wake-up times, estimated from their activity logs for the week of the study. We more fully describe how the final residual-score model was attained in a later section.

Rule Plot comparisons

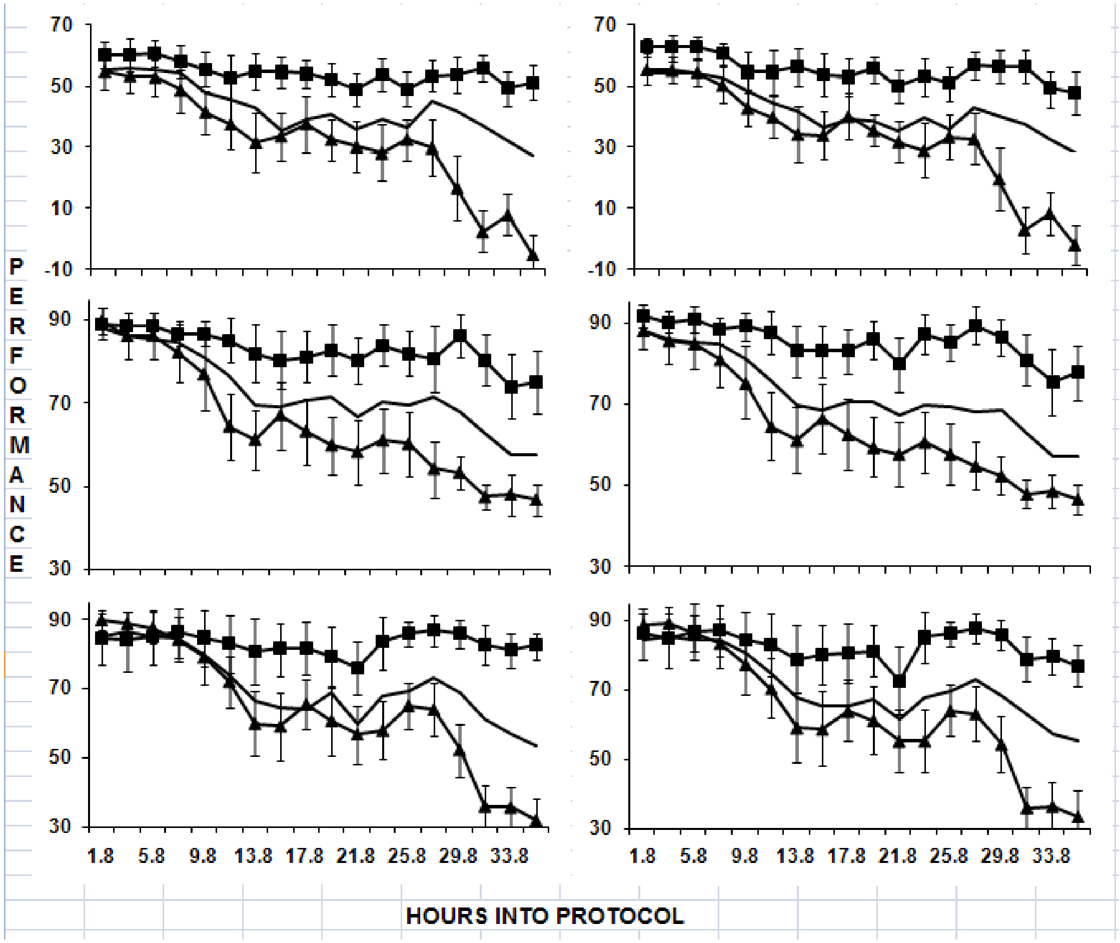
Figure 3 plots the percent-change rule and the residual-score rules, based on ANAM tasks, for the 3 possible groups of participants one can consider, namely the two extremes and the unclassifiable middle. Figure 4 plots the simple-change rule and the residual-score rule based exclusively on the PVT. As can be seen the two different rule-options provide similar plots with adequate separation between extreme groups, as indicated by standard error of the means, derived for each point.

Figures 3 through 4 suggest that the most fatigue susceptible groups have fatigue-impact functions most similar to SAFTE predictions. For example, the Continuous Performance Task in the bottom panel of Figure 3 shows the double-humped pattern in the plateau region most clearly in the susceptible group. That pattern is also visible in the aggregate tendency of performance functions in Figure 2. In SAFTE, the double-hump in the plateau region is expected as the result of two distinct circadian rhythms of different periods (Hursh, et. al. 2004).

Figure 3 Caption. Fatigue plots of raw data (performance scales as in Figure 1) broken down by fatigue classifications. Fatigue-resistant participants are square symbol, non-extreme middle has no symbol, and fatigue susceptible participants are a triangle symbol. The percent-change rule is displayed in the left column and the residual-score rule in the right column. ANAM tasks Math, Grammatical Reasoning, and Continuous Performance, are in rows 1, 2, and 3, respectively. To reduce clutter, error bars (as in Figure 1) are shown for susceptible and resistant groups only.

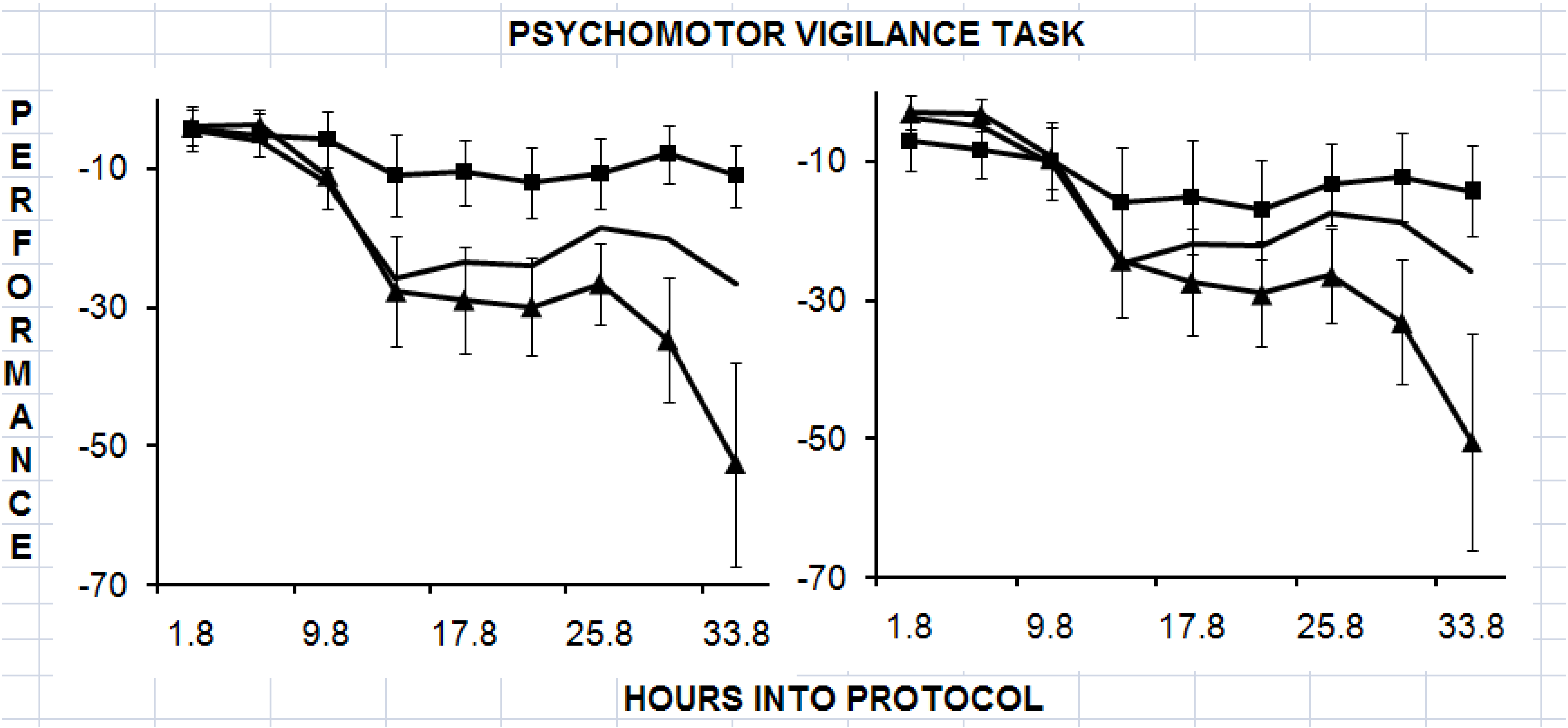
Figure 4 Caption. Fatigue plots of raw PVT data based on classifications via a simple-change rule (left-side) and via a PVT residual-score rule (right-side). Symbol, data, and error bars as described in Figure 3.

FIGURE 3



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

FIGURE 4



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

Rule Overlap comparisons

Overlap between any two rules can be assessed by examining 2 x 2 cross-tabulations between their classifications. After we explain what we mean by this and define some terms of reference using an example, we will describe some results of selected rule comparisons. When results for ANAM-type rules are given, without corresponding PVT-type results, the PVT results would be similar.

First we assessed how similar the ANAM-based percent-change rule classified participants relative to the ANAM-based residual-score rule. We found 17 people classified as susceptible by both rules and 15 people classified as resistant by both rules. The overlap implied by this is relative to perfect agreement, which is not strictly possible given the percent-change rule has more resistors than the residual-score rule (i.e., 21 as compared to 20, owing to ties at the classification boundary). However, there were 20 observed fatigue resistors in the smaller rule that the larger rule could have agreed on, and this defines a limit of overlap. There is also a "disagreement" diagonal, in which one rule calls a participant susceptible and the other rule calls the same participant resistant; however, ANAM percent-change and residual-score rules never contradict each other this way. Finally, there is another sense of "rule disagreement," which is more apparent in the size of the overlap numbers. This disagreement involves one rule classifying a participant as susceptible or resistant, while the other rule puts the same participant in the middle. We prefer to think about this as "rule-uniqueness" rather than disagreement. In the ANAM percent-change and residual-score comparison, there are eight classifications that are unique to each rule (i.e., 3 in the susceptible category and 5 in the resistant category).

As a related comparison, we considered the residual-score rule, without any sleep behavior correction -- i.e., *just* controlling for baseline performance and nothing else -- and compared this to the percent-change rule. We found 19 agreements on susceptibility, 18 agreements on resistance, and again no disagreements. The higher overlap for the basic residual-score and percent change rules shows that it is mainly the "controlling" for sleep behavior, and not the type of the rule, which increases our final residual-score rule's uniqueness relative to the percent-change rule.

Another comparison of interest is between rules that vary in the tasks they use, i.e., ANAM vs. PVT, respectively. For the percent-change rules, we found 10/20 agreements on susceptibles, 12/20 on resistors, and no disagreements. For the same type of comparison on residual-score rules (with full sleep-behavior correction), 11/20 on susceptible, 11/20 on resistors, but also 2 rule disagreements (one for each type). We might have suspected uniqueness for rules based on the PVT alone and other cognitive tasks from the factor-analytic result of Von Dongen et al. 2004, showing PVT and other cognitive tasks loaded different factors.[3]

As we have explored different metrics with the PVT, another comparison of interest is a classification rule based on PVT(errors) vs. PVT(rate). For the residual-type rule, we found the rules based on different PVT metrics have 10 agreements on susceptibles, 12 agreements on resistors, and two disagreements (one of each kind). In other words, eighteen out of the forty possible selections were unique to each rule, or about the same amount of rule uniqueness as implied by rules based on different tasks (i.e., PVT vs. ANAM).

Finally, we considered the rule overlap between the same rules but with different choices for what is considered the “fatigue endpoint”. That is, we very deliberately went out to 48 hours in our (estimated) sustained wake to define a fatigue endpoint. What would be the rule classification overlap between that choice and another hypothetical study that elected to end earlier than 48 hours? This would seem to have some practical importance for designing future studies, because a very high overlap would mean one could make do with less than a 48-hour fatigue endpoint. To make this comparison we considered the 4 testing trials just prior to the four terminal trials. This corresponds to a participant release time of about 2100 on the night before our actual 0600 release time (or an estimated sustained wake of about 39 hours). For the percent-change rule, we note 12/20 agreements on susceptible classifications and 11/21 agreements on the resistor classifications, and one disagreement between rules. For the residual-score rule, we note 13/20 and 14/20, susceptible and resistor agreement with one disagreement. The somewhat higher overlap for the residual-score rule may reflect the effects of correction, i.e., participants have more equated sleep characteristics, such as sleep and wake times the week of the study, with residual-score type rules, than they do for the percent-change comparison (as we will demonstrate later in Table 8).

Fatigue correlations between tasks

Both percent-change and residual-score rules grade participants on a range of fatigue-susceptibility. This allowed us to assess correlations of “fatigue effects” between tasks for both percent change and residual scores. As a safeguard against outliers we rank fatigue scores before correlating. The results are shown in Table 6. As can be seen, the size of the correlations is sometimes large, which imply fatigue impact is reliable across tasks. As these are difference scores, or change scores, or scores constructed by subtracting rested variability from fatigued variability, the size of the correlations is also large relative to a long-standing differential literature that finds difference scores to be less reliable than the simpler components going into them (e.g., as when one condition is subtracted from another and used as a score, c.f., Lohman, 1994).

Table 6

Percent-change ranks correlated between tasks (upper triangular matrix) and residual-score ranks correlated between tasks (lower triangular matrix)

	1.	2.	3.	4.	5.	6.	7.
1. Math		.56	.85	.58	.67	.43	.47
2. Gram	.55	--	.71	.48	.59	.22 ^{ns}	.57
3. CPT	.80	.67	--	.70	.69	.42	.59
4. PVT(errors)	.54	.40	.59	--	.50	.23 ^{ns}	.69
5. SYNWIN	.66	.61	.68	.45	--	.34 †	.45
6. C3STARS_Ind	.40	.43	.46	.37 †	.35 †	--	.11 ^{ns}
7. PVT(rate)	.51	.54	.59	.66	.45	.11 ^{ns}	--

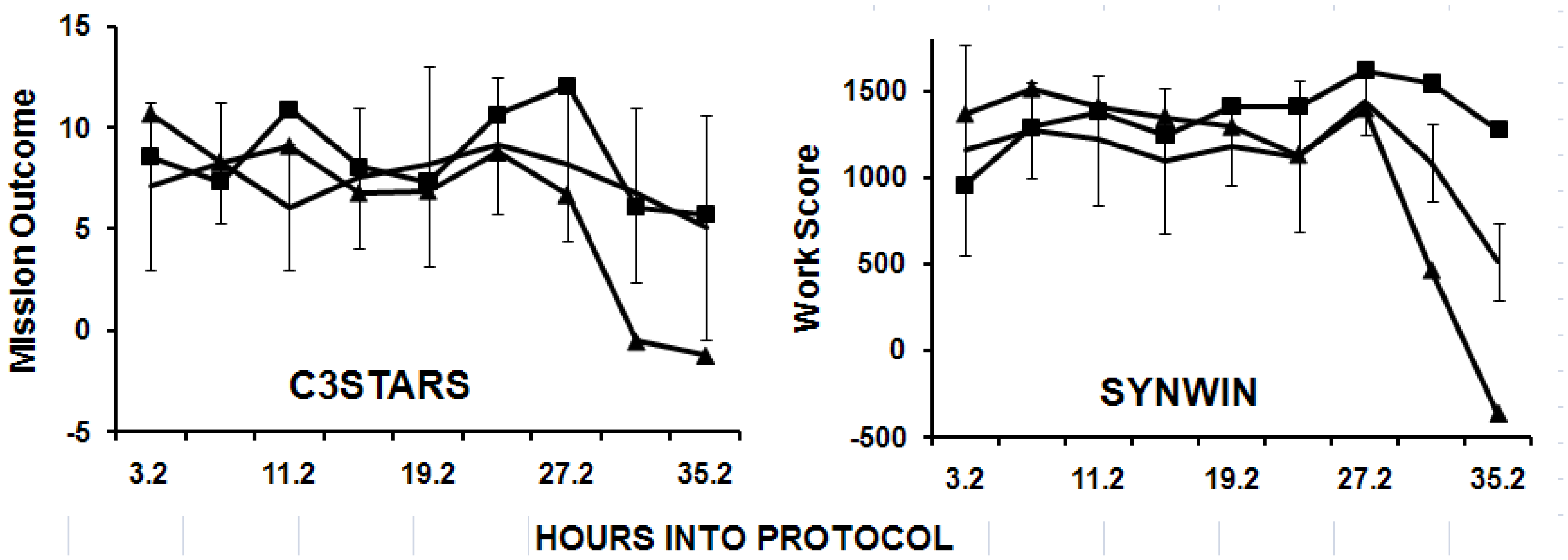
Note. Same notes as Table 5.

Fatigue classification impact on "simulated work"

Figure 5 plots simulated work performance on C3STARS and SYNWIN, for each participant classification based on the ANAM-type residual-score rule. This gives additional information from just viewing the correlations in Table 6. In particular, we note that the form of the fatigue functions for simulated work (when averaged over our participant classifications) would be different from the functions displayed in Figure 1. For instance, if one excludes the last 4 trials of C3STARS and the last 2 trials of SYNWIN, the main effect of fatigue is either not significant for C3STARS (*F*(13, 585) = 1.44) or significant but with cell means more consistent with learning for SYNWIN (*F*(6,522) = 6.43, *p* < .001). Although these work-like tasks are not as well trained as the ANAM, so that learning can and does continue to occur during the fatigue protocol, the fact that learning may be happening up to 39 hours in a sustained wake seems worth noting (c.f. Whitmore, et. al. 2007). We take up possible explanations (including learning) for these differences between work-like and cognitive tasks in another venue. For now, we note that our classification methods show the expected generalization to work-like tasks; albeit only at the fatigue endpoint region of the protocol.

Figure 5 Caption. Fatigue plots of raw data for simulated work for fatigue resistant (square symbol), non-extreme middle (no symbol), and fatigue susceptible (triangle symbol) groups, based on the ANAM residual-score rule. To reduce clutter error bars (in the manner of Figure 1) are shown only on the middle function. For C3STARS only even trials are plotted, as these are closest in time to SYNWIN administrations.

FIGURE 5



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

Demographics

Table 7 shows the demographics and sleep behavior characteristics for our sample of 89 protocol completers. Weekday and Weekend Sleep were significantly different: $t(88) = -7.25$, $p < .0001$ and uncorrelated ($r(89) = .16$), suggesting our participants use the weekend for recovery sleep and do not get their preferred amount of sleep during the week. The variable Recent Sleep is total sleep reported 2 nights before the study (going further back on the activity logs, would have reduced our *n* owing to missing data). Using Recent Sleep as a sample estimate of habitual Weekday Sleep (i.e., divide Recent Sleep by 2), we found Recent Sleep to be larger than reported Weekday Sleep ($t(83) = -3.27$; $p < .002$, mean hours 7.51 vs. 6.94, respectively, and these two variables were uncorrelated ($r(84) =.10$). The mean difference suggests participants slept about a half-hour more per night than usual, just prior to the study.

Table 7

Demographic Statistics

	Mean	Std Dev	Min	Max	N
Age	26.8	5.0	19	39	89
Female	.40	.49	0	1	89
Education	2.6	.85	1	5	86
Military	.64	.48	0	1	89
Tobacco Use	.16	.37	0	1	89
Video Game Enjoyment	6.6	3.0	0	10	61
Weekday Sleep	6.9	1.0	4	9.5	89
Weekend Sleep	8.2	1.4	5	11	89
Recent Sleep	15.0	2.7	9	22	84
Recent Wake Time	7.0	1.6	3.8	12.3	89
Recent Sleep Time	23.7	1.7	20.8	6.2	89

Notes. Female, Military, and Tobacco Use are binary variables scored 1 or 0. Education: 1- 5 (1 is High-school and 5 is beyond Bachelors, 2.6 reflects finishing high school and having some college). Video Game Enjoyment: rating of 1 to 10 on "how much do you enjoy video games?" Weekday Sleep: hours of sleep per night "on average" during the weekdays .Weekend Sleep: hours of sleep per night "on average" during weekends. Recent Sleep: cumulative hours of sleep reported for Weds and Thurs prior to the study (Friday). Recent Wake Time, Recent Sleep Time: average wake up and sleep time (24 hour clock, decimal fractions) reported for the week of the study. Sample size on demographics varies due to missing data.

Demographic correlation on task performance

Demographic to *ranked* task performance was investigated in an exploratory fashion, adopting an *alpha* .05, two-tail criterion. Correlations were observed for both baseline and fatigue-endpoint components of our rule classifications.

For ranked baseline performance, we found the following. Age was negatively correlated to C3STARS and SYNWIN (*r*(46) = -.42, *p* < .003; r(88) = -.25, p < .02, respectively). Being female and having a higher level of education was negatively correlated with C3STARS performance (*r* (46) = -.35, *p*<.02; *r*(45) = -.35, *p* < .02, respectively). Saying one enjoyed video games went with higher performance on C3STARS, SYNWIN, and ANAM Continous Performance Task (*r*(33) = .61, *p* < .001; r(60) = .40, p < .001, r(61) = .27, *p* < .04, respectively). Tobacco went with lower Grammatical Reasoning performance (*r*(89) = -.21, *p* < .05). More Weekday Sleep went with better initial performance on the Continuous Performance Task (*r*(89) = .28, *p* < .01) and Recent Sleep went with better performance on PVT(errors) (*r*(84) = .27, *p* < .02) and SYNWIN performance (*r*(83) = .22, *p*<.05). Finally, Recent Wake Time went with better initial performance on SYNWIN (*r*(88) = .23, *p*<.03).

For ranked fatigue-endpoint performance correlated to factors in Table 7, we found fewer significant relationships. However, an important finding was that Weekday Sleep correlated more to fatigue endpoint on ANAM tasks than to baseline performance on the ANAM tasks, *r*(89) = .26, *p* < .01; *r*(89) = .24, *p* < .03; *r*(89) =.34, *p* <.001 for Math, Grammatical Reasoning, and Continuous Performance Task, respectively. PVT(errors) at fatigue-endpoint was a near-miss (*r*(86) = .21, p < .056). Habitual Weekday Sleep relates to chronic sleep deficit in our population sample; and therefore, its effects might be more apparent after a sustained wake than at the initial performance baseline. Additionally, Recent Wake Time also relates to fatigue endpoint performance (later wake times going with better performance), though not always to our stated level of significance (for ANAM Math, Grammatical Reasoning, and Continuous Performance Tasks: $r(89) = .21$, $p < .06$, $r(89) = .21$, $p < .06$; $r(89) = .37$, $p < .001$, respectively; for PVT(errors) $r(86) = -.26$, $p < .02$; for SYNWIN $r(89) = .42$, $p < .001$).

We might have predicted PVT fatigue-endpoint to be more sensitive to chronic sleep debt (i.e., Weekday Sleep) than the ANAM measures. Lower reliability for the PVT from fewer trials estimating its fatigue-endpoint (i.e., 2 trials as opposed to 4 trials for the ANAM) is the likely cause. However, lack of sensitivity for the PVT cannot be attributed to our metric choice for the PVT. We assessed other PVT metrics against Weekday Sleep. As with other correlations assessed, we transformed every metric to ranks. PVT(rate) did not correlate higher than our chosen metric (i.e., r(86) = .18, $p < .09$).

As another kind of comparison between metrics, we considered how well PVT simple-change predicted aggregated ANAM percent-change (i.e., that variable we used to define classification for the percent-change rule). Our chosen metric, PVT(errors), did better than PVT(rate), though not significantly ($r(86) = .66$, $p<.001$ as vs. $r(86) = .60$, $p < .001$, respectively).

Demographic correlation to classification rules

We finally get to our method of correcting the percent-change rule from the effects of irrelevant (i.e., non-genetic factors) that may bias a participant's classification. We correlated the classifications of the ANAM percent-change rule (a binary variable, susceptible vs. resistant) to the demographics/sleep behavior factors of Table 7, and we did the same for the PVT simple-change rule. Demographics that were significantly correlated (i.e., $p < .05$, two-tailed) to either the ANAM or PVT *uncorrected* change-based classifications were considered in residual scoring to correct for the bias.

For the ANAM percent-change rule, the first six (demographic) factors of Table 7 showed no significant relations (range of *r*s -.12 to .15). For the last 5 (sleep behavior) factors, Weekday Sleep correlated $r(41) = .36$, $p < .02$ (means: 6.6 vs. 7.4 hours for susceptible and resistant groups, respectively) as did Recent Wake Time $r(41) = .43$, $p < .005$ (means: 0610 vs. 0737, susceptible vs. resistant groups, respectively). We can also look at these correlations over the full range of 89 participants and over a more extreme classification (i.e., top and bottom 11 of the 89 participants) as a gauge of the stability of the relations. For the full sample, we found $r(89) = .29$, $p < .007$ and r $(89) = .36$, $p < .001$, for Weekday Sleep and Recent Wake Time, respectively. For the more extreme classification rule, we found $r(22) = .47$, $p < .03$ and r(22) = .37, $p < .09$. Hence, the relationship between Weekday Sleep and the percent-change rule is robust by being present on the entire sample and on two different definitions for our extreme groups. Recent Wake is perhaps less robust, but nevertheless the relation between it and the percent-change rule for ANAM is significant in the full sample.

For the PVT simple-change rule, Tobacco Use was positively related to resistance ($r(41) = .31$, $p < .05$). Though participants were not allowed to use Tobacco during the protocol, 5 out of the 6 Tobacco Users that were classified by the simple-change rule were found to be in the resistant group as defined by PVT(errors). Recent Wake and Sleep Times were correlated to PVT classification, i.e., resistors woke up about an hour later than susceptibles ($r(41) = .39$, $p < .02$) and went to sleep about 50 minutes later than susceptibles ($r(41) = .31$, $p < .05$).

Given these findings we elected to use Weekday Sleep, Recent Wake Time, and Recent Sleep Time as additional covariates in the creation of residual-scores for the participants, thus allowing a ranking that is an alternative to raw percent change. Relative to the SAFTE model these are the rational factors to correct for, i.e., habitual sleep amounts and circadian phase are the major determinants of cognitive effectiveness in that model. While total Recent Sleep on the week of the study doesn't correlate significantly to classification, it is implicitly coded in the bias-correction model from having an individual's average going-to-sleep times and waking-up times in the correction model. That is, total Recent Sleep the week of the study can be estimated by the former covariates, and the estimate correlates strongly (*r*(82).=.81) to Recent Sleep.

We elected *not* to correct PVT for Tobacco Use as literature links nicotine dependence to genetic causes (Haberstick, et al., 2007). We are on somewhat shaky ground by including Wake and Sleep Time estimates as correction factors in our bias model. If considered as indices for morningness/eveningness (i.e., circadian *preferences* in our participants) there is an ambivalent literature supporting genetic determinants of that (see the literature cited by Von Dongen, Vitallero, & Dinges, 2005, pg. 481). On the other hand, we have models of fatigue that say that when you wake up can impact how well you do by the end of the protocol (i.e., Figure 2 contrasts of two SAFTE predictions), so we assumed it was legitimate to control for recent sleep behavior.

Ranking the residual scores yields different participant selections for our extreme groups than ranking participants on raw percent-change as the earlier section on rule overlap showed. Table 8 shows mean differences on the factors of the bias model, both for the residual score classifications and the uncorrected percent-change classifications. The participants in the extreme groups defined via the residual-score rule do not *significantly* differ on the sleep characteristics corrected for; however, they can still have trends in the bad direction with respect to the corrected characteristic (e.g., Weekday Sleep, still has a 20 min per night difference between resistant and susceptible groups, t(40)=1.44). This may be a “penalty” for the number of items corrected for, given when we correct only for Weekday sleep (and no other factor), mean differences on *Weekday Sleep* between Resistant and Susceptible groups are less apparent (e.g., on that way of defining the residual-score rule, the resistant group would sleep slightly *less* than the susceptible group).

Table 8

Mean differences between susceptible and resistant extreme groups on classification bias factors for the Percent-Change/Simple-Change rules (which are uncorrected for the bias factors) and the Residual-Score rules (which are corrected for the factors)

	ANAM				PVT			
	Percent-Change		Residual-Score		Simple-Change		Residual-Score	
	delta	t-value	delta	t-value	delta	t-value	delta	t-value
Weekday Sleep	.73	2.42*	.38	1.44	.36	1.2	.22	.70
Recent Sleep	.62	1.46	.13	.51	.81	2.06*	.10	.13
Recent Wake	1.45	2.99*	.01	0.3	1.1	2.64*	.29	.50

Note. *Delta is the mean difference between susceptible and resistant groups on the row correction factor. All deltas are in hour units. t-value is for the between-groups t-test with n=21,20 on uncorrected rules and n = 20,20 on residual-score rules.*

** p < .05, two-tailed.*

DISCUSSION

We found that individual-differences in cognitive abilities, as indexed by a set of tasks, increase with amount of sustained-wake or fatigue. This goes against the notion that the impact of fatigue on cognitive systems destroys individual-differences in cognitive abilities (e.g., flooring everybody). A substantial number (e.g., 20) of flat-functioned or fatigue-resistant individuals could be found, replicating Von Dongen, et al's. (2004) finding of a trait-like characteristic for fatigue-resistance observed over multiple sleep deprivation sessions. In our particular case, that observation is within a single session. An additional empirical observation we can make, given our greater exploration of the raw data under our classifications (i.e., Figures 3 through 5) is the apparent insensitivity of initial baseline performance on a task (i.e., aptitude for the task) for predicting who will be susceptible or resistant to fatigue. In only one task and under one rule, was group membership significantly related to how well one did at the beginning of the protocol (i.e., the ANAM Math, for the residual-score rule, 61.2 vs. 66.5, $t(38) = 2.33$, $p < .025$, 2-tailed, susceptible vs. resistant, respectively). Therefore, when we say individual-differences are increasing under fatigue, we are not saying that the "cognitively-rich" are getting richer and the "cognitively-poor" are getting poorer, but that some are sustaining and some are not, at all levels of initial performance.

Von Dongen, et al.'s procedures are clearly superior to ours regarding knowing the initial condition of participants prior to sleep deprivation, as they rigorously controlled participants' sleep behavior in the laboratory for 3 weeks prior to the repeated measurement. When sleep history is not so rigorously controlled, as in our study, we found, from sleep behavior questionnaires and activity logs, potential biases in our participants' classifications. We made the assumption that these biases were non-genetic, or more related to life-style choices. We then tried to correct for these biases, given follow-on correlational analyses to genetic data were being planned. We found extending a residual-score (to account for the biases) attenuated bias relative to an uncorrected percent-change rule (Table 8). The approach of considering ranks of residual-scores is more efficient than assessing participants on raw percent-change and then manually matching people on one or more potentially confounding factors, as this matching procedure substantially reduced our *n* for follow-on analyses.

We also note that the residual-score created *just* from using initial performance showed very high overlap to the percent-change rule with regard to classifications. We think percent-change and the most basic residual-score look alike because our experimental conditions *approximate* the condition of our subjects not varying (much) on baseline performance but substantially varying on endpoint performance. Under such conditions, simple residual and percent-change scores are both functions of the endpoint scores and constants, and therefore should rank participants similarly. These conditions would also lead to reliable difference scores as evidenced by cross-task correlation of fatigue impact across tasks (Table 6).

However, correction via residual scores comes with a substantial problem, which is a variant of the "correlation is not causation" problem. In the case of sleep history, there are strong theories that expect certain factors to bias our classifications, but in the case where there are no such theories, correction can be too much of a judgment call. As an example suppose Video Game Enjoyment *did* significantly relate to a PVT classification rule. Should that demographic be co-varied in an attempt to index a person's fatigue resistance more accurately?

We don’t know. As the demographic may cause the performance (as in a bias) or be a common expression of a genetic trait that causes both the performance and the demographic, correcting may or may not be removing a genetic trait of interest.

Fatigue results related to a fatigue theory (part 2)

Given our characterization of SAFTE as a task-invariant model, how can it provide help in the identification of fatigue-resistant and susceptible individuals? One interesting possibility is parsimony in a bias-correction model. Had we known our participants’ sleep histories 30 days out from the beginning of their sustained wake, we could have generated a tailored SAFTE prediction for each participant for fatigue-endpoint performance (or for any point of the protocol). While such wouldn’t be a performance estimate for the task, per se, the “cognitive effectiveness” for fatigue endpoint, which is strictly determined by sleep history, could have been used as a single covariate to handle, en masse, all the sleep history covariates we considered in our (ad hoc) bias-correction model.

To be validated for this kind of eventual use, the model (e.g., SAFTE) would have to be subject to iterative improvements, where deviations of observed data from the model’s expectation could be explained, the explanations tested, and the model improved. However, there is a long-standing issue for how to assess the goodness of fit of task-dependent performance data to a *task-invariant* fatigue model. To make such a comparison, one often sees observed fatigued performance divided by baseline performance (e.g., Hursh, et. al. 2004, Figure 6), or a percent-change transformation on observed data. But how appropriate is this transformation for comparing to model predictions? As a basis for doubt, we note dividing by a task’s baseline score yields functions that depend on the specifics of baseline scale (i.e., an RT metric is not guaranteed to fatigue, percent-change wise, in the manner of an accuracy metric, even on the same task).

In our study, we compared a task's average observed intra-subject z-score (hereafter ISZ) to a similar transform of the SAFTE predictions. This is a strategy, which is different from the concept of percent change, for putting different kinds of numeric series (e.g., performance data and "physiological" prediction) on the same scale metric (see Von Dongen, 2004, for a different procedure). To see how *z*-transforming data and prediction is relevant to "model fitting," we first give the reader a more precise sense of what the ISZ transformation, or any z-transform, does.

We state as a theorem (provable with some simple algebra), that if two number series are mutually reachable via linear transformation, then they *must* have the same *z*-transform, or *z*-transforms that are negatives of each other (as in two lines with positive and negative slope). Therefore, a series with scores {1, 20, 1, 20, 1} has the same *z*-transform as {100, 101, 100, 101, 100}, and a series with scores {2, 4, 6, 8, 10} has the same *z*-transform as {0, 500, 1000, 1500, 2000}. As these series pairs could represent participant scores on some task, one can see, z-transforming, and the ISZ in particular, is *very inappropriate* for comparing participants to each other, so why is *z*-transforming useful at all?

When we convert both a SAFTE prediction and an observed fatigue function via a *z*-transform, we are reasoning the following way. If the predictions *did* reach the observed data by a linear transformation, then the predictions would fit the data in that sense. Conversely, if they cannot, there is something worth explaining in the data that the theory cannot. We consider the act of *z*-transforming both theory and data a checking operation that tests the explicit hypothesis "similar by virtue of a linear transformation," and we can use error bars, in an analogous sense to more canonical model fitting, to inform on the likeliness of any deviations of observed from predicted. This characterization of "fitting the data" puts no restriction on the form of the model expectations, as *z*-transforming only "tests" whether model output and empirical "output" look the same, in the sense of being just a linear transformation away from each other. The benefit of accepting this point of view as a kind of "model-fitting" is that the deviations between PVT metrics, errors and rate, and between observed test performance and model expectation, become interesting enough to try to explain as we do so now.

We observed that PVT(rate) and PVT(errors) produced different fatigue functions under ISZ, but the difference was localized to the beginning of the protocol where fatigue impact should be least. This observation, coupled with the observation that the two PVT metrics correlate least when fatigue is least, suggest PVT(rate) is more of a chimera of two functions within an individual, the first part being related to motor speed and the second part reflecting fatigue impact. If so, the larger *F* for this metric in Table 3 could be explained by extra range being added to the performance where there is no fatigue. This is a radically different explanation than say, PVT(rate) being more sensitive to fatigue owing to its larger *F* for the trial effect. This kind of counter argument can only be made by observing the ISZ functions of the different metrics. A percent-change metric *will not detect this*, given that metric constrains all tasks to start out at 100% (although an ISZ of percent-change would still show the divergence, as the ISZ of raw-data and the ISZ of percent-change are the same, owing to percent change being a linear transformation of raw data).

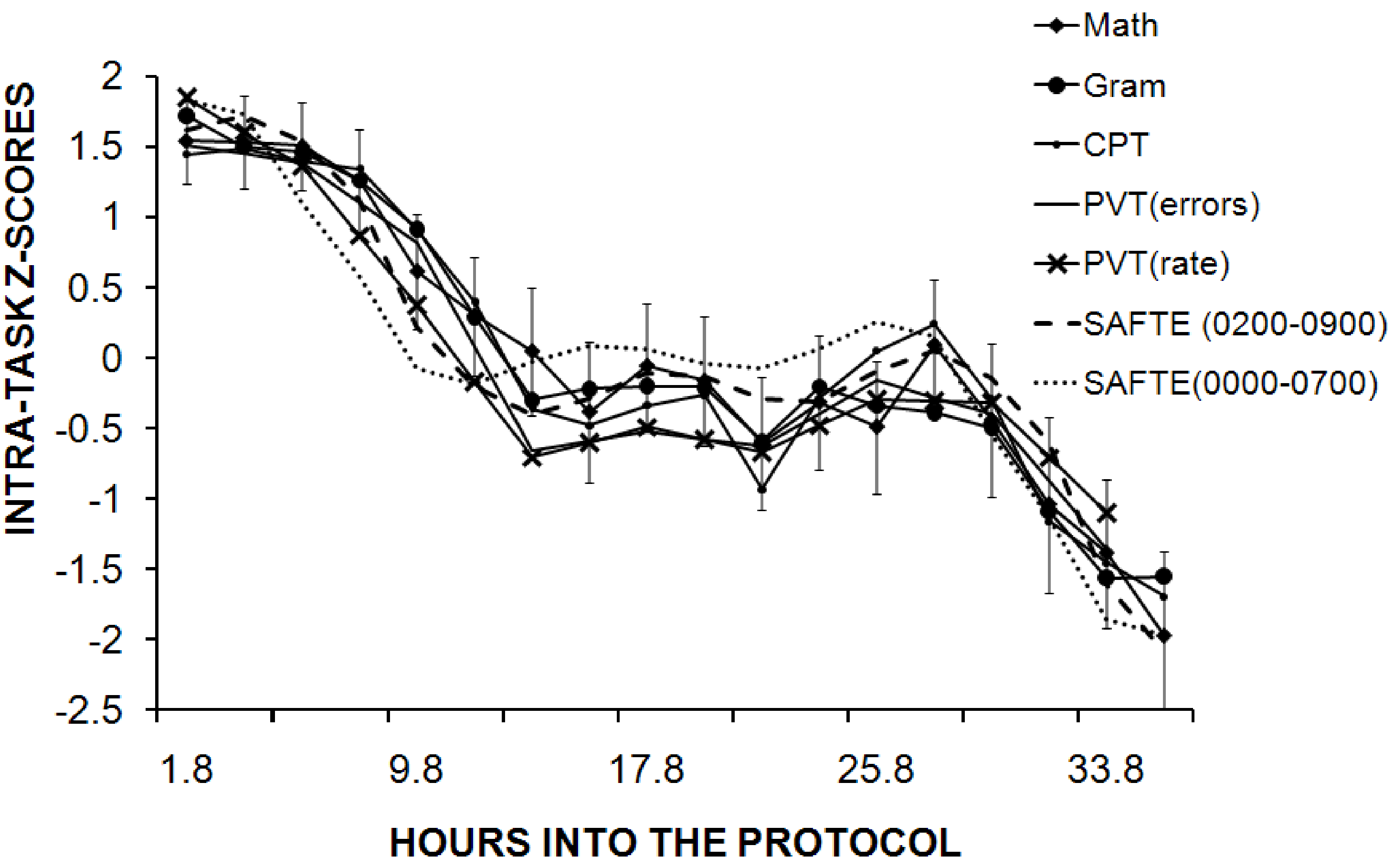
The ISZ also took the ANAM tasks and PVT(errors), which are tests with different cognitive contents and different scales of measurement, and expressed performance on them in such a way that similar fatigue impacts were seen on these tasks. Why didn't SAFTE predictions follow the tasks, or why was the range of model prediction, in ISZ, greater than the observed range of ISZ? One might be tempted to blame the tasks for showing a floor effect on the expression of fatigue, although it is not obvious that the ANAM tasks are at floor, given omitted responses are counted wrong. Another possibility is to blame the model, which is the same as blaming the participants for not acting like the model. In this regard, the individual differences in fatigue-resistance may be tempting to cite as a contributing factor to the systematic lack of fit in Figure 2. However, being insensitive to fatigue does not necessarily mean model "atheoretic." In our discussion of the ISZ above, we showed very differently sloped functions (which could reflect different participants with different fatigue sensitivity) could also have the same ISZ. Finally, one can blame an aspect of the application of the *z*-transform as being inappropriate for assessing SAFTE model fits. Given SAFTE predictions are derived from model parameters that were estimated by *mean* data from other studies (Hursh, et. al., 2004); this is the most likely explanation.

In fact, if SAFTE were not a model that had been optimized to predict *mean* data, the use we described above for it (i.e., a parsimonious representation of sleep history bias for susceptible/resistant classifications) would not make sense. If SAFTE did not predict *mean* performance, then a residual obtained from predicting a person's observed performance from their SAFTE cognitive effectiveness estimate *would not be* a measure of under or over-achievement with respect to the model's expectations. Mean data, of course, contain the variance of *between-subject* and *subject x fatigue* interactions, whereas ISZ does not contain the former variance, and contains the latter variance to a diminished degree.

However, that last criticism is specific to the ISZ, not the z-transform, as a quick and dirty model-fitting procedure. To see this, Figure 6 is a recasting of Figure 2; however, the z-transform is now applied to average task performance, which is the result of the variance components ISZ either ignores or attenuates. That is, Figure 6 treats each task like a "participant," whose observed scores are the mean of scores derived from the 89 participants (i.e., what is plotted in Figure 1), so we might call this use of the z-transform the "Intra-Task" z-score. In Figure 6 the SAFTE predictions are the same as Figure 2, as is the scale. Figure 6 does not show the large *systematic* lack of fit for SAFTE that is apparent in Figure 2. We attribute this to the more appropriate assessment of observed *mean* data, which is what SAFTE has been engineered to predict.[4]

Figure 6 Caption. Fatigue plots, attained from z-transforms of average performance functions for multiple tasks compared to SAFTE fatigue model predictions (which are on the same scale as Figure 2). Error bars are derived from the standard errors of the means in the raw scale. All else is the same as in Figure 2.

FIGURE 6



Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.

One can also see that the variability of the intra-task z-plots is larger than the ISZ plots (also as expected), so it may not be too clear that the absolute fit relative to the *best* SAFTE prediction is getting better in Figure 6. However, the sum-squared deviations of the observed functions (i.e., ANAM and PVT(errors)) from the best SAFTE prediction (i.e., the 0200-0900 schedules) is larger in Figure 2 than it is in Figure 6 (i.e., 12.895 vs. 6.051, respectively). When this ratio of the sum-squared deviations is tested, as a difference between two variance estimates, the test would be modestly significant ($F(62,62) = 2.13$, $p < .05$), indicating better fit for SAFTE in Figure 6 (i.e., variability around the best prediction line is smaller in Figure 6).

Conclusions: fatigue models and fatigue data

The ISZ transformation was not appropriate for assessing a model fit of SAFTE, but was useful in other sorts of tests, owing to its ability to represent abstract information about the shape of a function. Given this, ISZ could test whether two metrics on the same task had the same or different shape with respect to their fatigue functions (e.g., PVT(errors) and PVT(rate)). ISZ could also show how different tasks with different scales and content had similar sensitivity to fatigue. We have also used the ISZ to assess whether *task conditions,* such as, team vs. individual C3STARS performance, could be shown to be differentially sensitive to fatigue, in a more conservative fashion than assessments on the raw data alone (in preparation).

There was also evidence that the SAFTE model fit our observed data when assessed on the intra-task *z*-scale. However, we have to take this last finding with a grain of salt, as we really don't *know* that the SAFTE profile we assessed for fit (0200-0900) is the correct one (and if our recent sleep logs are more accurate as estimates of habitual sleep behavior, we suspect it is not). We have to leave the definitive answer to that question to future studies executed toward that goal. However, our larger point is that the assessment could be made using the intra-task z. Given this, we wonder why the transformation is not more often used in the literature to assess fatigue models.

REFERENCES

Bonnet, M. H. (2000). Sleep deprivation. In M. Kryger, T. Roth, & W. Dement (Eds.), *Principles and practices of sleep medicine* (pp. 53-68). Philadelphia, PA: W. B. Saunders Company.

Caldwell, J. A., Mu, Q., Smith, J. K., Mishory, A., Caldwell, J. L., Peters, G., Brown, D. L., George, M. S. (2005). Are Individual Differences in Fatigue Vulnerability Related to Baseline Differences in Cortical Activation? *Behavioral Neuroscience,* 119(3), 694–707.

Chee, M. W. L, & Choo, W. C. (2004). Functional Imaging of Working Memory after 24 Hr of Total Sleep Deprivation. The Journal of Neuroscience, 24(19), 4560–4567

Dinges D., Pack F., Williams K., Gillen K.A., Powell J.W., Ott G.E., Aptowicz C., & Pack A.I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep,* 20, 267–77.

Dorrian, J., Roach, G. D., Fletcher A., Dawson, D. (2007). Simulated train driving: Fatigue, self-awareness and cognitive disengagement. Applied Ergonomics, 38, 155–166.

Elsmore, T. F. (1994). SYNWORK: A PC-Based Tool for Assessment of Performance in a Simulated Work Environment. *Behavior Research Methods, Instrumentation, and Computers* 26, 421-426.

Galliaud, E., Taillard, J., Sagaspe, P., Valtat, C., Bioulac, B., & Philip, P. (2008). Sharp and sleepy: evidence for dissociation between sleep pressure and nocturnal performance. *Journal of Sleep Research,* 17, 11–15.

Gunzelmann, G., Gluck, K. A., Kershner, J., Van Dongen, H. P. A., & Dinges, D. F. (2007). Understanding decrements in knowledge access resulting from increased fatigue. In The 29th Annual Conference of the Cognitive Science Society. Nashville, Tennessee, USA.

Haberstick, B. C., Timberlake, D., Ehringer, M.A., Lessem, J. M., Hopfer, C. J., Smolen, A., & Hewitt, J.K. (2007). Genes, time to first cigarette and nicotine dependence in a general population sample of young adults, *Addiction*, 102, 655–665.

Harrison, Y., & Horne, J. A. (2000). The impact of sleep deprivation on decision making: A review. *Journal of Experimental Psychology: Applied*, 6, 236-249.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D.R., Belenky, G., Balkin, T.J., Storm, W. F., Miller, J. C., & Eddy, D. R. (2004) Fatigue Models for Applied Research in Warfighting. Aviation Space and Environmental Medicine, 75(3), 44-53.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! Intelligence, 14, 389-434.

Kryger, M., Roth, T., & Dement, W. (2000). *Principles and practices of sleep medicine* (3rd ed.). Philadelphia, PA: W. B. Saunders Company.

LeDuc, P.A., Caldwell, J. A., & Ruyak, P. S. (2000). The effects of exercise as a countermeasure for fatigue in sleep-deprived aviators. *Military Psychology*, 2000, 12(4), 249–266.

Lohman, D.F. (1994). Component scores as residual variation: or Why the intercept correlates best. *Intelligence*, 19, 1-11.

Melissa M. Mallis, M.M., Mejdal, S, Nguyen, T.T., and Dinges D. F. (2004). Summary of the key features of seven biomathematical models of human fatigue and performance. *Aviation, Space, and Environmental Medicine*, 75(3), A4-A14.

National Institute for Occupational Safety and Health. (2004). *Overtime and Extended Work Shifts* (Report No. 2004-143). Cincinnati, OH: Author.

National Sleep Foundation. (2005). *Sleep in America poll*. Washington, DC: Author.

Olofsen, E., Dinges, D.F., & Van Dongen H.P.A (2004). Non-linear mixed-effects modeling: individualization and prediction. *Aviation, Space, and Environmental Medicine,* 75(3), A134-A140.

Pilcher, J. J., & Huffcutt, A. I. (1996). Effects of sleep deprivation on performance: A meta-analysis. *Sleep*, 19, 318-326.

Reeves, D., Winter, K., Kane, R., Elsmore, T., & Bleiberg, J. (2001). *ANAM 2001 user's manual* (Special Report NCRF-SR-2001-1). San Diego, CA: National Cognitive Recovery Foundation.

Tessier, P. (2006) Command, Control, and Communications, Simulation, Training, and Research System (C3STARS). Unpublished computer program.

Van Dongen, H. P. A. (2006). Shiftwork and inter-individual differences in sleep and sleepiness. *Chronobiology International*, 23(6): 1139–1147.

Van Dongen, H.P.A., Maislin, G., & Dinges, D.F. (2004). Dealing with inter-individual differences in the temporal dynamics of fatigue and performance: importance and techniques. *Aviation, Space, and Environmental Medicine*, 75(3), 147-154.

Van Dongen, H.P.A, Baynard, M.D., Maislin, G., Dinges, D.F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: evidence of trait-like differential vulnerability. *Sleep*, 27, 3, 423-433.

Van Dongen, H.P.A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviation, Space, and Environmental Medicine*, 75(3), A15 - A36.

Whitmore, J., Chaiken, S. R., Harrison, R., Harville, D. (2007) Sleep loss in complex team performance. In Waard, D. D., Hockey G. R. J., Nickel P., Brookhuis, K. A. (Eds.), *Human factors of performance in complex systems*, Shaker Publishing, Maastricht, The Netherlands.

Whitmore, J., Doan, B., Fischer, J., French, J. & Heintz, T. (2004). The Efficacy of Modafinil as an Operational Fatigue Countermeasure Over Several Days of Reduced Sleep During a Simulated Escape and Evasion Scenario. AFRL-BR-TR- 2004-021.

Woltz, D.J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General*, 117, 319-33

Author Note

This research was conducted by the Biobehavioral Performance Branch under work unit 7757P907 and 7757P918. We thank Phil Tessier, Teresa Cassiano, and Tammy Savoie for substantial work in the execution of this study. The opinions expressed in this article are the authors and do not necessarily reflect those of the Air Force. Order of authorship for the first two authors was established via a coin flip. This research significantly extends and refines preliminary results reported in the first part of AFRL-RH-BR-TR-2008-0050. Correspondence should be sent to Scott Chaiken (scott.chaiken@brooks.af.mil) or Don Harville (donald.harville@us.af.mil).

Footnotes

[1] On odd ANAM trials only, Time Stream 1 has about 10 mins of mood scales given after ANAM and PVT, on the ANAM computers; whereas the same mood scales were given before the ANAM and PVT in Time Stream 2. As we don't analyze this data, we don't show the task in the testing events, although the effect of their being there helps equate (odd-trialed) ANAM testing times across the two streams.

[2] In fact, CPT has a more accentuated circadian, or a significant time by *z*-score interaction against Math or Gram considered separately; whereas ANAM Math and Gram considered against each other do not show significant differences.

[3] We also did exploratory factor analyses on just the 3 ANAM tasks, PVT(errors), and SYNWIN (5 variables n=86). We found that using a minimum eigenvalue of 1.0 all tests loaded a single factor. However, when we requested 2 factors (varimax rotation), PVT was the high marker on the second factor, whether we considered corrected residual scores, raw fatigue-endpoint, or raw initial performance scores (second factor eigenvalues, .83, .82, and .95, respectively).

[4] Hursh, et al. 2004 did not model mean data but *mean percent-change* data. We note that in our data the difference between an intra-task z based on averages of raw data (which is what we did) correlates above .99 to intra-task *zs* based on averages of percent-change transformed data (e.g., for ANAM tasks and PVT(rate)). In any case, our argument is not about what

function is *best* to z-transform, but about supporting the Intra-task-*z,* a transform on mean data, as more appropriate for assessing SAFTE model fit than the ISZ.

APPENDIX: C3STARS DESCRIPTIONS

Figure A1 Caption. This figure shows task-specific events and their approximate times in a scenario with time running down the figure. Within a row activities in boxes are parallel. Also past a certain point (shown as a dashed line) phases overlap, so both rows and columns of boxes indicate parallel activity

Figure A2 caption. This figure shows a screen dump of C3STARS from the solo-player condition (middle lane being played). A bomber (B23) and a fighter (FE4) have SAM and hostile air missions respectively. A mission is indicated by a commit line and color (e.g. red color intercept hostile, violet color rendezvous with friendly). ISR (IH3, IL4, IL3) assets are shown with their tracking rings (blue cyan lines) which are related to whether ordinary or BDA information can be gathered on a shot SAM. Note IL3 is hidden by hostile fighter Z503. SAM icons show varying amount of identification (e.g. wavy line indicating probable, e.g. SA6H, probable decoy, and SA3E, as unknown). Red and Blue arcs in front of asset show ordinance range. Significant events are issued in text in the communication window and delayed auditory (given speech is turned on), e.g. FE4 has seen its target but it is outside missile range ("FE4 Bandit"). Switch actions and simulation controls, in the form of buttons are above the situation display.

.

FIGURE A1

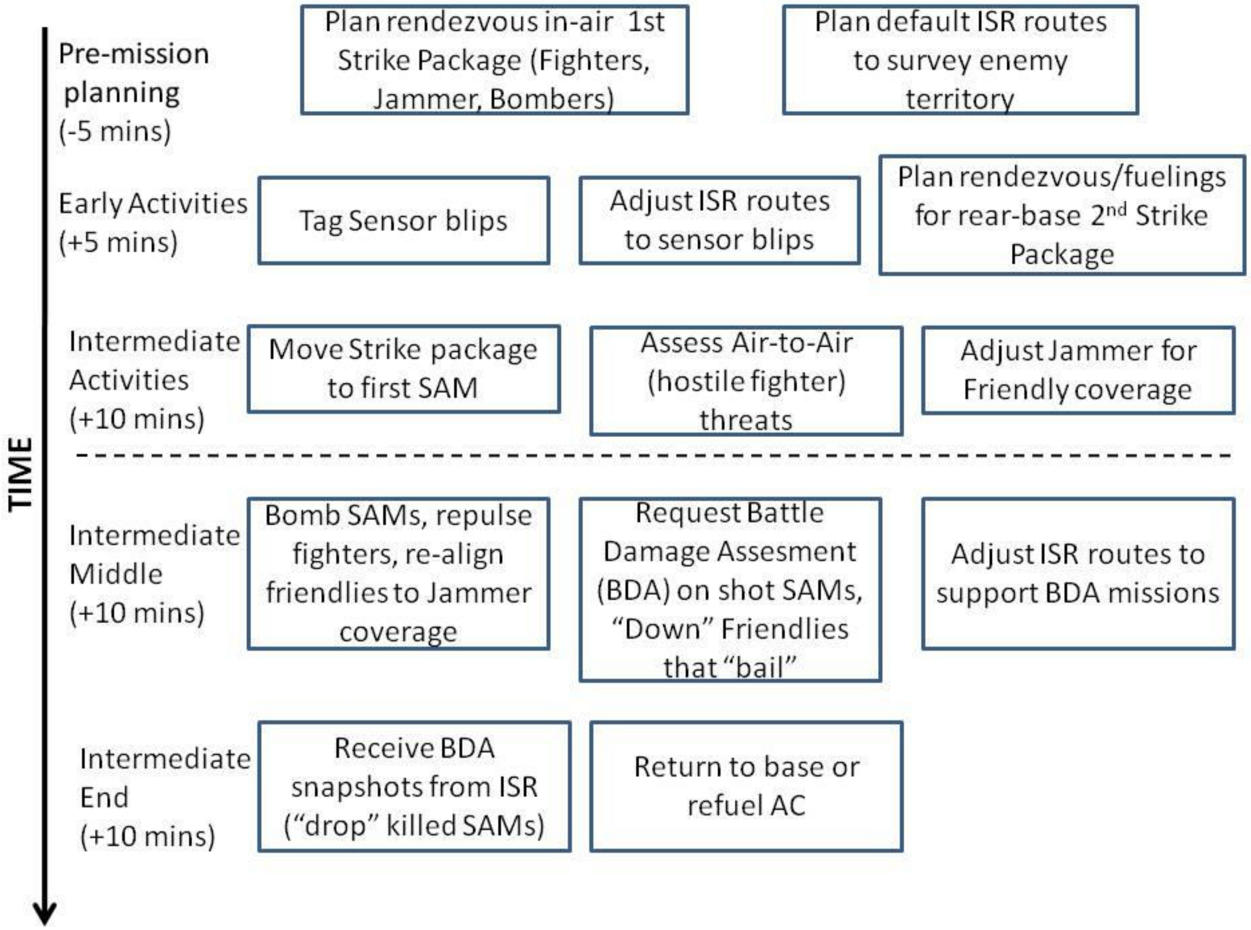
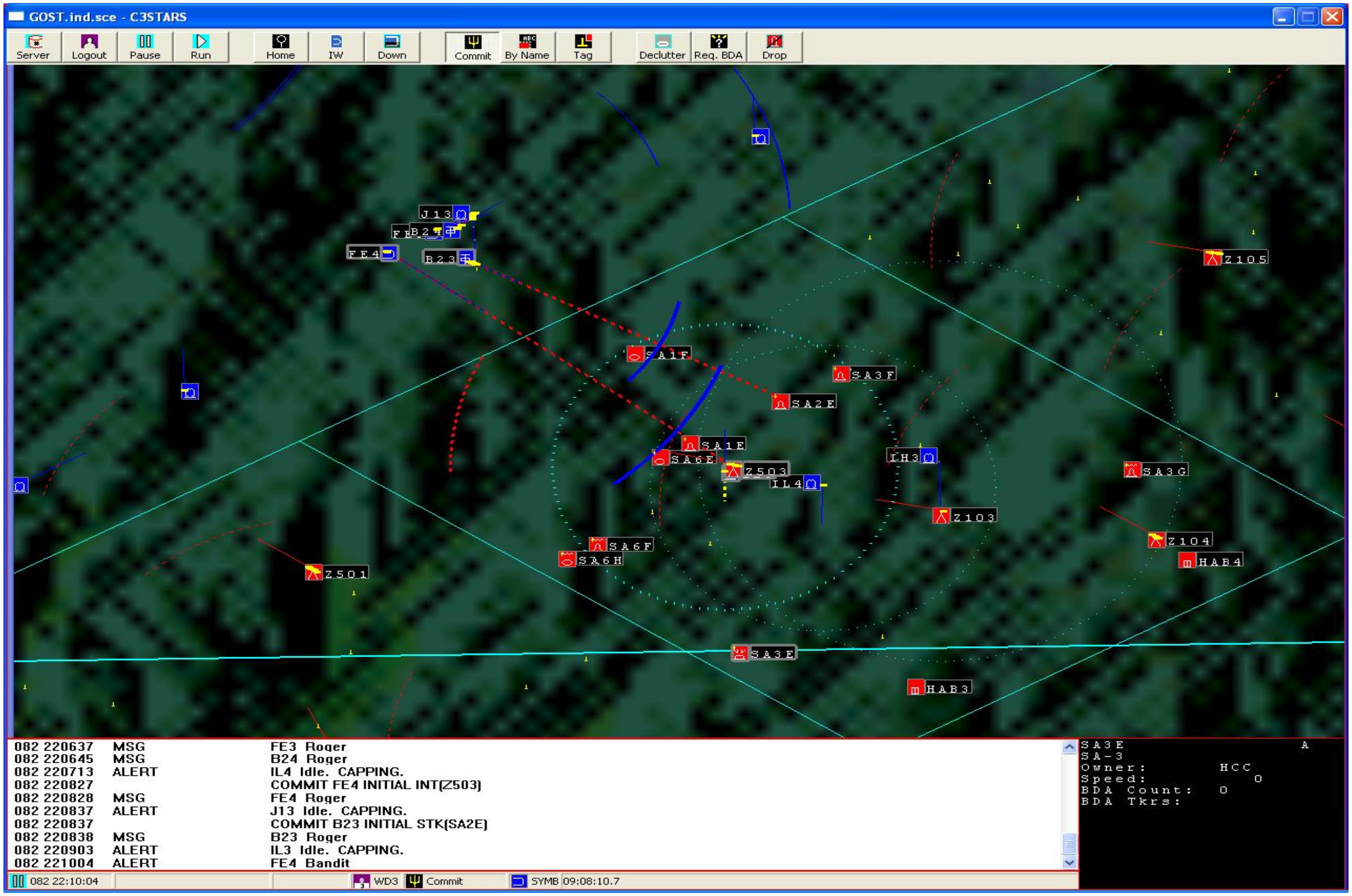

TIME
Pre-mission planning (-5 mins)
Plan rendezvous in-air 1st Strike Package (Fighters, Jammer, Bombers)
Plan default ISR routes to survey enemy territory
Early Activities (+5 mins)
Tag Sensor blips
Adjust ISR routes to sensor blips
Plan rendezvous/fuelings for rear-base 2nd Strike Package
Intermediate Activities (+10 mins)
Move Strike package to first SAM
Assess Air-to-Air (hostile fighter) threats
Adjust Jammer for Friendly coverage
Intermediate Middle (+10 mins)
Bomb SAMs, repulse fighters, re-align friendlies to Jammer coverage
Request Battle Damage Assesment (BDA) on shot SAMs, "Down" Friendlies that "bail"
Adjust ISR routes to support BDA missions
Intermediate End (+10 mins)
Receive BDA snapshots from ISR ("drop" killed SAMs)
Return to base or refuel AC

FIGURE A2



```
082 220637   MSG       FE3  Roger
082 220645   MSG       B24  Roger
082 220713   ALERT     IL4  Idle.  CAPPING.
082 220827             COMMIT FE4 INITIAL INT(Z503)
082 220828   MSG       FE4  Roger
082 220837   ALERT     J13  Idle.  CAPPING.
082 220837             COMMIT B23 INITIAL STK(SA2E)
082 220838   MSG       B23  Roger
082 220903   ALERT     IL3  Idle.  CAPPING.
082 221004   ALERT     FE4  Bandit
```

```
SA3E                A
SA-3
Owner:      HCC
Speed:          0
BDA Count:  0
BDA Tkrs:
```

082 22:10:04 | WD3 | Commit | SYMB 09:08:10.7

Approved for public release; Distribution unlimited, Case File No. 09-038, January 2009.